The Philosophy of Science in a European Perspective

Maria Carla Galavotti
Dennis Dieks
Wenceslao J. Gonzalez
Stephan Hartmann
Thomas Uebel
Marcel Weber *Editors*

New Directions in the Philosophy of Science

Springer

New Directions in the Philosophy of Science

[The Philosophy of Science in a European Perspective, Vol. 5]

Proceedings of the ESF Research Networking Programme

THE PHILOSOPHY OF SCIENCE IN A EUROPEAN PERSPECTIVE

Volume 5

www.pse-esf.org

For further volumes:
http://www.springer.com/series/8745

Maria Carla Galavotti • Dennis Dieks
Wenceslao J. Gonzalez • Stephan Hartmann
Thomas Uebel • Marcel Weber
Editors

New Directions in the Philosophy of Science

Springer

Editors
Maria Carla Galavotti
Department of Philosophy
and Communication
University of Bologna
Bologna, Italy

Wenceslao J. Gonzalez
Faculty of Humanities
University of A Coruña
Ferrol, Spain

Thomas Uebel
School of Social Sciences
University of Manchester
Manchester, UK

Dennis Dieks
Institute for History and Foundations
of Science
Utrecht University
Utrecht, The Netherlands

Stephan Hartmann
Center for Mathematical Philosophy
Ludwig Maximilian University
of Munich
Munich, Germany

Marcel Weber
Department of Philosophy
University of Geneva
Geneva, Switzerland

ISBN 978-3-319-04381-4 ISBN 978-3-319-04382-1 (eBook)
DOI 10.1007/978-3-319-04382-1
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014940727

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume, the fifth in the series *The Philosophy of Science in a European Perspective*, collects selected articles from presentations delivered at the three events organised in 2012 by the European Science Foundation Research Networking Programme PSE (The Philosophy of Science in a European Perspective): (1) the conference "New directions in the philosophy of science" held on October 17–20 at the Bertinoro Conference Centre of the University of Bologna; (2) the workshop "Causation, dispositions and probabilities in physics and biology" that took place on November 22–24 at the University of Lausanne, and (3) the workshop "Philosophy and the sciences – old visions, new directions" held on November 30–December 1 at the University of Cambridge, on the premises of CRASSH (Centre for Research in the Arts, Social Sciences and Humanities).

The Bertinoro conference resulted from the synergy of the five teams of researchers belonging to PSE, namely: Team A: "Formal methods" (leader Stephan Hartmann, co-leader Thomas Müller); Team B: "Philosophy of the natural and life sciences" (leader Marcel Weber, co-leader Hanne Andersen); Team C: "Philosophy of the cultural and social sciences" (leader Wenceslao J. Gonzalez, co-leader Amparo Gomez); Team D: "Philosophy of the physical sciences" (leader Dennis Dieks, co-leader Guido Bacciagaluppi); and Team E: "History of the philosophy of science" (leader Thomas Uebel, co-leader Michael Stoeltzner). Each of these teams organised one main session and one junior session, all revolving around the central topic that imprinted the research carried out by PSE in its fifth year of activity, namely "New directions in the philosophy of science".

The Lausanne workshop originated from a project of Michael Esfeld, member of PSE's Steering Committee, in close cooperation with the leaders of Teams B and D. The papers read there aimed at investigating possible links between biology and physics in connection with the notions of causality and dispositions, taken in a probabilistic fashion. While such notions play an important role in biology, it is unclear whether the same holds for physics. It turns out that focussing on these notions can shed light on still unexplored relations between these two major fields of research in the natural sciences.

The Cambridge workshop linked the newly-established CamPoS (Cambridge Philosophy of Science) research group to PSE, and was locally organised by Huw Price in collaboration with PSE's chairperson Maria Carla Galavotti. The workshop focussed on the relationship between Cambridge and Vienna in twentieth century philosophy of science, with the hope that this relationship will again come to play a major role in European and world philosophy of science in the twenty-first century. Six mini-symposia, each hosting two speakers, were held at the workshop, plus two junior sessions comprising four papers each.

Since all three events pointed in some way or other to new trends in the philosophy of science, with special emphasis on research carried out in Europe, it was decided to arrange the contributions collected in this volume in five sections, corresponding to the five PSE teams, irrespective of whether they were delivered in Bertinoro, Lausanne or Cambridge. However, it does not seem out of place to recall to which of the three conferences they originally belonged. The names of the authors are listed here in the order in which their contributions appear in this volume. The Bertinoro conference hosted the papers of Thomas Müller, Liesbeth De Mol, Patrick Suppes, Raffaella Campaner, Jeroen Van Bouwel, C. Kenneth Waters, Pierre-Luc Germain, Wolfgang Spohn, Matti Sintonen, Daniel Andler, Tarja Knuuttila, David-Hillel Ruben, Katarzyna Paprzycka, Obdulia Torres González, Chiara Ambrosio, Christopher A. Fuchs, Guido Bacciagaluppi, F.A. Muller, Miklós Rédei, Michał Marczyk and Leszek Wroński, Pablo Acuña, Ronnie Hermens, Petr Švarný, Huw Price, Massimo Ferrari, Thomas Uebel, Matthias Neuber, Uljana Feest, Sean Crawford, Anastasios Brenner, and Cristina Chimisso. The Lausanne workshop hosted Mark Colyvan, Tim Räz, Jan Faye, Jan Baedke, Max Urchs, Raphael Scholl, Cristian Saborido, Andreas Bartels and Daniel Wohlfarth, Mario Hubert and Roland Poellinger, Claus Beisbart, Radin Dardashti, Luke Glynn, Karim Thébault, Mathias Frisch, Gábor Hofer-Szabó, Dustin Lazarovici, Tomasz Placek, and Dennis Dieks. The Cambridge workshop hosted the papers of Kerry McKenzie, Veli-Pekka Parkkinen, Tim Lewens, Maria Carla Galavotti, Henrik Rydenfelt, and Friedrich Stadler.

This volume ideally represents PSE's point of arrival after five years of activity starting in 2008. The other volumes in the same series are: *The Present Situation in the Philosophy of Science* (proceedings of the conference held in Vienna, 18–20 December 2008), edited by Friedrich Stadler, Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Thomas Uebel and Marcel Weber, published in 2010; *Explanation, Prediction, and Confirmation* (proceedings of the workshops held in 2009), edited by Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Thomas Uebel and Marcel Weber, published in 2011; *Probabilities, Laws, and Structures* (proceedings of the workshops held in 2010), edited by Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Michael Stoeltzner and Marcel Weber, published in 2012; and *New Challenges to Philosophy of Science* (proceedings of the activities held in 2011), edited by Hanne Andersen, Dennis Dieks, Wenceslao J. Gonzalez, Thomas Uebel and Gregory Wheeler, published in 2013.

Having directed the ESF programme PSE from beginning to end, and as the principal editor of this volume, I am proud to say that together with the others

the present volume reflects the vitality and originality of European philosophy of science. It is widely recognised that during the last five years the world scenario of philosophy of science has become more balanced, with a significant number of research groups and important events taking place in Europe. Without a doubt, PSE's activities and publications played a major role in this development. On behalf of the European community of philosophers of science, I wish to express our deep gratitude to the European Science Foundation for having supported our research in this field.

Bologna, Italy Maria Carla Galavotti

Contents

Part I
Formal Methods

Things in Possible Experiments: Case-Intensional Logic as a Framework for Tracing Things from Case to Case

Thomas Müller

1 Introduction

Science needs modality. Science is about finding out what the world is like and what there is – but it is also about finding out what the world could be like and what there might be. While this may be controversial when taken as a metaphysically loaded claim about some ultimate picture of reality, it is just a simple descriptive truth when one takes actually practiced science into account. That practice is often not just about writing down what happened where and when, but about studying the things involved in such happenings, and finding out what could possibly happen with them. A large amount of the vocabulary used in the sciences is dispositional in nature, and while this may be more easily visible in the so-called special sciences like biology, it is also true of fundamental physics. One does not even need to focus on linguistic issues to see the importance of modality. Think of experiment, a crucial ingredient in modern science: experiments consist in the manipulation of the course of nature in the interest of scientific insight – in active intervention on what is happening, in order to bring about something else. The very notion of an experiment presupposes an acknowledgment of different possible courses of events. Experiments also involve the prevention of unwanted disturbances – vibrations, electrical fields, or variations of temperature, depending on the case – and shielding these off can be difficult and costly; so it is important to know which disturbances are possible and what their effect would be.

T. Müller (✉)
Department of Philosophy, Utrecht University, Janskerkhof 13a, 3512 BL Utrecht, The Netherlands

Fachbereich Philosophie, University of Konstanz, Fach 17, 78457 Konstanz, Germany
e-mail: Thomas.Mueller@phil.uu.nl; Thomas.Mueller@uni-konstanz.de

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_1,

Given that we have a need for modal notions in science, it is interesting to ask which kinds of modality are involved, and how we can best understand them. In this paper we will focus on one aspect of these questions: we will look at the *representation of things in modal contexts* occurring in science. We proceed from the assumption, not to be argued for here, that formal methods of philosophical logic are often helpful in elucidating philosophical problems arising in philosophy of science, and consequently we approach our topic from a logical point of view. We will, however, keep our formalism minimal.

The paper is structured as follows. We begin by describing the use of possible experiments in science and in everyday life (Sect. 2). This will make salient the notion of tracing a thing from a given case to other possible cases. In Sect. 3 we describe this notion in a more formal way, exhibiting some consequences for standard systems of quantified modal logic. In Sect. 4, we introduce CIFOL, case-intensional first order logic, as a newly established formal framework that helps to elucidate the notion of tracing. We wrap up in Sect. 5.

2 Possible Experiments

It seems best to start with an example to introduce our general topic. We will look at possible experiments that could be conducted to find out an object's trait, while we ascribe that trait in any case. We start with an object's charge. Charge is a quantitative property of microscopic as well as ordinary objects. Much of the chemical behavior of atoms is explained by their being made up of charged particles, and when you carry a sufficient amount of charge, it makes your hair stand on end. Charge appears to be quite a fundamental property – it is surely objective, mind-independent and categorical if anything is. Charge does not appear to be a disposition of an object like fragility or solubility. But still, it has modal force, and that may even be one of its defining features.

Consider Coulomb's law, which links the force F acting on a test charge of q situated at a distance r from an object, to the charge Q of that object:

$$F = \frac{1}{4\pi\epsilon_0}\frac{Qq}{r^2}. \tag{1}$$

There are various ways to read this equation. One sensible way to read it is as a criterion of when the object we are interested in, let us call it α, has a charge of Q. Read in that way, the equation explains our epistemic access to an object's charge via an experiment, and we actually conduct such experiments in some cases: if we put a test charge of known charge q at a distance r from α, we can use Coulomb's law to determine α's charge. (Of course, in an actual application of this recipe, we will need to make sure there are no unwanted interferences – that is what laboratories are for. Read as a law, (1) has the familiar feature of holding only *ceteris paribus*. Also,

there will have to be some way of determining the resulting force F from *its* effect, e.g., via the acceleration of a particle of known mass. We will leave these additional complications out of consideration.)

It seems sensible to say that for any charged object α, there should be some way of conducting an experiment of the mentioned type that will reveal α's charge. (Call this "modal verificationism" if you like. We will not argue for this doctrine's universal validity here, which would probably get us entangled in problems involving masking and similar effects. Never mind *that* for the time being – it seems sensible to work out the base case first.) Now, what is the link between a situation before us, in which we have identified an object α that we are interested in, and the mentioned merely possible experiment? It seems reasonable to spell out this link in terms of a counterfactual: if α, the thing before us, were tested in an experiment, the outcome would be such-and-such. (Again, leaving well-known problems aside.) In discussions of this approach, the focus is mostly on identifying the correct counterfactual case (or set of counterfactual cases) that needs to be considered. Certainly it will have to be one in which, contrary to what is happening to α in the case at hand, α is being subjected to an experiment of the mentioned type; the question then is what *else* also needs to change. (For example, I am in fact now writing this paper in a café. In a counterfactual situation in which I am now conducting the experiment, it must be assumed that I went to the lab instead of going to the café.)

In the identification of such a counterfactual case, one often invokes a notion of similarity of cases (or "possible worlds"), e.g., via a similarity ordering among cases. While many interesting issues are involved in developing the notion of similarity, we will not follow this line of investigation here. Rather, we will treat the more basic question of what we mean by re-identifying the object α, which we can identify directly (e.g., by pointing) in the case at hand, in some other possible case. It is certainly crucial that in another case such as the one involving the mentioned experiment, we keep our focus on the object, α, and do not switch to something else, since it is α's charge that we are interested in. It is also crucial that in following α, we treat it as a charged body, so that we incorporate equality of charge somehow into the adequacy condition for reidentification, and thereby make sure that we do not end up with an object that we may identify as α in *some* sense, but which will not do for identifying α's current charge.

Here is a more concrete example that brings out what is at stake. Consider my daughter's cat, Hannibal. I am interested in finding out his present mass, since I suspect that he is becoming a little chubby lately. Finding out his mass involves subjecting him to a sort of experiment, e.g., putting him on scales.[1] Now if it's really his exact *present* mass that I am interested in (suppose that I bet with my

[1]Mach's definition of mass via a possible experiment, which does not involve a detour via an object's weight in a gravitational field, is analyzed by Bressan (1972), whose formal investigations are a main source of inspiration for the work presented here – see Sect. 4. Mach's definition is based on an experiment in which the object under consideration bumps into the unit mass at some specified speed. My daughter surely won't let *that* happen to her cat.

daughter that he is over 6 kg, and precision is required), I need to make sure that his mass does not change between the present case and a case in which he is put on the scales. Hannibal's present mass is not simply he mass of the cat, Hannibal, in a case in which he is put on scales. He was actually put on scales when he was a kitten, and weighed less than 2 kg then, but this obviously contributes nothing towards an answer to our question. Even starting out now, he might be put on scales after he has had some extra food, or shed some of his fur or otherwise lost some weight, which would lead to an incorrect answer. In order to get the right answer, it will not do to just follow the cat between the case at hand and some case with the cat on scales. Rather, I need to make sure that I am following the *massive object* that is identical to the cat right now, and subject *it* to the weighing experiment. What needs to be traced is the cat as a massive object, not the cat as a biological individual, since the property under consideration, mass, is one of massive objects and only derivatively one of biological individuals.

It is the same with charge. When we see the cat's hair stand on end and become interested in his current charge, it won't do to take him to the lab and bring some test particle close to him – chances are that in that case, his charge will already be quite different from what it was in the situation that is of interest. Some more elaborate scheme of tracing the charged object that is identical to the cat before me will have to be found.

Note that all of this does not mean that properties of the cat are "really nothing but" properties of the cat's matter. First, even in identifying the cat's mass or charge *now*, we do not need to trace the cat's current matter – that would be highly impractical, since a living cat constantly exchanges matter with his environment, and that matter quickly disperses all over the place, e.g., by diffusion. In the case of mass, it will be enough if we can trace the living cat and just make sure that the input/output mass balance is neutral,[2] and similarly for charge. Second, if we are interested in whether the cat has the property of responding to his name, we arguably need to trace the biological individual, the cat, to a case in which the appropriate experiment is conducted, i.e., in which he is called and responds – or doesn't.

It seems, then, that for various scientific and everyday purposes, it is important to understand how an object identified in a case at hand can be traced to other cases in which that object is involved in a possible experiment. In the remainder of this paper, we will look at formal means for the representation of such tracing.

3 Tracing in Standard Quantified Modal Logic

We have argued for the importance of the notion of tracing a thing between possible cases, or reidentifying a thing in another, merely possible case. From the formal-logical point of view that we are taking in this paper, this notion of tracing should be

[2]This will obviously involve keeping the cat away from his food; depending on the precision required, ambient humidity and other factors may also have to be taken into account.

connected with the handling of variables and terms in a predicate logical framework that also allows one to talk about different possibilities. That is, we are entering the realm of quantified modal logic.

The construction of systems of quantified modal logic was approached syntactically by Barcan (1946), and later also semantically, most notably by Kripke (1959). Despite several refinements and much discussion, that latter framework can still serve to lay out the general idea behind most currently available systems of quantified modal logic. There are two main components. On the one hand, there is the handling of modality: the most common image is that of a set W of different possible worlds as modal alternatives – perhaps with an accessibility relation between them that grounds relational modal semantics, as in Kripke (1963). These possible worlds $w \in W$ provide a global view on modality; they correspond to complete alternative ways our world could be.[3] On the other hand, there is the handling of individuals: These are conceived of as inhabitants of the various possible worlds, so that each world w comes with its own domain of quantification, D_w. While there are important differences in the treatment of quantification in different systems, the common idea is that at a world w, a variable should have as its value some $d \in D_w$. (Arguments then arise, e.g., about the interrelation of the different D_w, or about the handling of reference failure.)

Based on this background, there are two main ideas for expressing the notion of tracing, or reidentifying, an object across different possible worlds. One idea, propounded by Kripke (1980), is that variables function as *rigid designators*, i.e., that they have the same value at each world. Another idea is to deny that different worlds can host the same individual, i.e., to deny so-called trans-world identity. On that approach, due to Lewis (1968), the worlds and domains need to be supplemented by a (perhaps context-dependent) *counterpart relation* that associates an inhabitant of a world w with its counterpart (or counterparts) in any given other world w'.[4] On both approaches, a metaphysical conviction settles details of the logical handling of variables.

In our view, both of these approaches are inadequate for capturing the phenomenon of tracing. Both either have a hard time handling the temporal aspect of modal alternatives in our examples, or make it difficult to represent the different, yet objectively grounded tracing principles of physics vs. biology that the examples point out.

The first problem is due to the image of possible worlds itself. In describing our examples, we were using the normal English idiom of possible cases, according to which necessity is truth in any case. Now "case" is not a technical term, and there are certainly useful applications of modal notions in which we consider world-spanning global possibilities, or possible worlds – e.g., when it comes to

[3] Lewis (1986) famously proclaims that according to his doctrine of modal realism, every such way our world could be, is a way some world actually *is*. The theory of possible worlds is, however, independent of that doctrine.

[4] See Kracht and Kutz (2007) for a detailed exposition of the formal aspects.

models of cosmology. But cosmology is not a typical science: it has to do without experiments and to rely on observation instead. In everyday life and in sciences such as chemistry, biology and many branches of physics, we are much more interested in what is possible locally rather than globally. This is to be expected if science appeals to experiment, since we, the experimenters, bring about effects in the world locally, manipulating one thing (or a few things) at a time.

So, it seems necessary to allow for a more local notion of modality when trying to capture the notion of tracing – we do not normally trace objects from possible world to possible world, but from (local) case to (local) case. Such tracing has a temporal aspect. But the notion of rigid designation makes little sense when temporal cases are allowed. Certainly it needs to be possible to represent things as changing – but then, what does it mean that the value of a variable stays the same through different temporal cases, as rigid designation would demand? It seems that the image of a thing behind this approach is that of a bare particular, a mere peg to hang properties on that is devoid of any structure. We are not saying that it is not possible to build a quantified modal logic with rigid designator variables that handles the issue of temporal cases in some way – e.g., by representing change exclusively on the side of properties. But such a move seems awkward at best.

Counterpart theory seems to be better suited for tracing changing objects over time. After all, the counterpart relation can be *anything*. But that is exactly the problem with that framework. We would like to have a logic that, while allowing some sensible variation in tracing principles (as in the cat/massive object example of Sect. 2), also embodies at least some clear formal constraints on what can be a counterpart of what when moving from case to case.

We are not claiming that these problems provide decisive arguments against attempts of building quantified modal logic on a quantification theory involving rigid designation or counterpart theory. What we will do in the following, is rather to sketch an alternative, such that those interested in the issue may judge for themselves. This alternative, case-intensional first order logic (Belnap and Müller 2013a), based on Bressan (1972), avoids taking sides in metaphysical issues such as trans-world identity, and instead offers a metaphysically neutral framework that, on our view, really helps to elucidate the notion of tracing a thing from case to case.

4 Tracing in CIFOL: Case-Intensional First Order Logic

We will now introduce our preferred logical framework for tracing things across possible cases: case-intensional first order logic (CIFOL), described in detail in Belnap and Müller (2013a).[5] As mentioned, that system takes its main inspiration from the work of Bressan (1972), an Italian physicist who developed his complex

[5]See also Belnap and Müller (2013b) for an extension to a system explicitly based on branching histories, which puts the general CIFOL machinery to work for a discussion of indeterminism.

modal-logical system ML^{ν} in order to better understand the modality involved in positing possible experiments in mechanics.[6] CIFOL is a first-order system of modal logic. Unlike Bressan's system, it does not allow for higher-order quantification over properties, relations or entities of higher types; its quantifiers are restricted to first-order entities, i.e., objects.

CIFOL's modality is based on a set Γ of cases, which do not have to be possible worlds, but may be temporal as well. For the basic system, nothing about Γ is assumed (except that it should not be trivial, i.e., it needs to have more than one member). Necessity is truth in all cases; formally:

$$\gamma \models \Box\phi \text{ iff for all } \gamma' \in \Gamma,\ \gamma' \models \phi.$$

Generally speaking, CIFOL is built on Carnap's method of extension and intension (Carnap 1947), which is applied universally to all parts of speech. Thus, each expression has an extension in each case, and an intension, which is the function from cases to the respective extensions. Formally:

$$ext_{\gamma}\xi = (int\ \xi)(\gamma); \quad int\ \xi = \lambda\gamma(ext_{\gamma}\xi).$$

For a sentence, ϕ, the extension in a case γ is a truth-value, **T** or **F**, so that instead of the "γ satisfies ϕ" locution, "$\gamma \models \phi$", we can also write "$ext_{\gamma}\phi = \mathbf{T}$". The intension of a sentence (a propositional intension) is then a function from the cases to truth values, representing the pattern of the extension of ϕ in case γ as γ is varied.

So much is standard in modal logic generally (even though it is often expressed differently). There are two features that set CIFOL apart and that allow for a useful analysis of tracing. First, the extension/intension method is applied to *all* terms, including variables (and even definite descriptions, but we will not go into that here). That is, a variable x has an intension, $int\ x$, and in each case γ, an extension, $ext_{\gamma}x$, that is a member of the extensional domain, D. Second, predication is generally intensional: whether a predication $\Theta(\alpha)$ is true or false in a case γ, need not be settled by the extension of α in γ alone. A predicate Θ therefore indicates, for each case γ, which *individual intensions* fall under it in that case. In contrast, standardly variables have just one constant extension (rigid designation), and predicates are extensional. There is another feature that strengthens CIFOL's expressive power: identity *is* extensional, i.e., whether two terms α and β are identical in a given case γ, only depends on these terms' extensions in that case, $ext_{\gamma}\alpha$ and $ext_{\gamma}\beta$. The semantic clause for identity statements is therefore:

$$\gamma \models \alpha = \beta \text{ iff } ext_{\gamma}\alpha = ext_{\gamma}\beta.$$

[6]The expressive resources of ML^{ν} are comparable to those of Montague's more well-known IL (Montague 1973). One important advantage of Bressan's system is uniformity: ML^{ν} does not require explicit type conversions, in contradistinction to IL.

These technical choices allow for an interpretation of the formalism that is helpful for understanding tracing. The main idea is to effect a *Gestalt* switch with respect to the intensions of terms (individual intensions). In standard quantified modal logic, the *extension* of a term in a case is taken to indicate which *object* is designated by the term in that case. In CIFOL, on the other hand, the idea is that the object designated by a term corresponds to the term's *intension*.[7] A term's *extension* in a case is therefore not an object; in fact, no metaphysical interpretation of the extensions is needed at all. It is enough if they are there to do the technical work of grounding the truth or falsity of identity statements.[8] It is therefore a simple and metaphysically innocent thing to say, e.g., that the *cat*, Hannibal, *is identical*, in the case before us, to a *lump of matter*, without that being necessarily so. In fact, if the cat is identical to some lump of matter in one case, he will not be identical to that lump of matter once he has taken another breath and thereby exchanged some of his matter with his environment. No fancy doctrine of contingent identity is needed to model what is going on here: we simply have (extensional) identity in one case, but not in other cases.

It should now be clear that the tracing of an object across possible cases in CIFOL is effected, quite simply, by the intension of the term, which represents the object. (Or: which represents the object's extension varying from case to case.) This is already enough to show how CIFOL overcomes the problems that are caused by taking variables to be rigid designators in standard quantified modal logic: in CIFOL, variables have intensions, which may be regular object-intensions, and a change in an object from case to case can simply be modeled by a change of the corresponding extensions from case to case.

Now it may seem that by allowing for more generality in the notion of a case and in the handling of variables, we have in fact succumbed to the doctrine of counterparts after all – isn't the variation of extensions from case to case pretty much the same as a counterpart relation between the extensions? If there were no further constraints on the intensions representing objects, the two systems would indeed show some similarity (even though there would still be the important dissimilarity that CIFOL extensions do not represent things). CIFOL, however, provides crucial extra resources to help limit allowable means of tracing in a formally lucid way. These extra resources come as non-creative definitions that can be formulated within CIFOL, not as changes to CIFOL's logical system. (In our view, this counts

[7] As variables are just terms, this interpretation accords with Quine's famous slogan that "to be is to be the value of a variable": the value of a variable is an intension, which represents an object.

[8] There is one special extension, denoted $*$, that is used to signal non-existence, as in failed definite descriptions; it is also useful to treat $*$ as a term always denoting $*$. In a temporal interpretation of CIFOL cases, it may be useful to think of the extensions as stages of objects, or as tropes. No matter which way; the important thing is *not* to think of these extensions as objects themselves. Technically, all that matters is the cardinality of the extensional domain D. It doesn't even matter whether the domains are taken to be case-relative or not. We work with the simpler choice of just one global extensional domain, D.

towards the usefulness of CIFOL as a *logic*.) The two main definitions are due to Bressan (1972), but CIFOL allows for some simplifications. The idea is to formulate constraints on those properties that can sensibly be taken to correspond to natural sortal properties (or substance sortals), by singling out certain first-order definable conditions on those properties.[9]

The first definition describes modal constancy: a property is *modally constant* iff an intension that falls under it in some case, falls under it in all cases. (Only a cat can be a cat; if something is possibly a cat, it is necessarily a cat.) This can be expressed as a condition on a predicate, Θ, as follows

$$\Box\forall x[\Diamond\Theta x \rightarrow \Box\Theta x].$$

The second condition is *modal separation*: it prescribes individual intensions falling under a sortal to be properly individuated, so that overlap of things of the same sort is precluded. (No two cats in exactly the same place.) Since we are dealing with sortals whose instances are contingent beings that can fail to exist in a case, we need to add a clause involving the existence predicate, E, so as not to forbid that two different things falling under the same sortal should fail to exist in the same case.[10] As a condition on a predicate, Θ, modal separation reads as follows:

$$\Box\forall x\forall y\,[(\Theta x \wedge \Theta y \wedge Ex \wedge Ey \wedge x = y) \rightarrow \Box(x = y)]$$

Putting these two definitions together, we can define a tracing property, or a *CIFOL sortal*, to be any predicate Θ that is both modally constant and modally separated.[11]

With all of this machinery in place, we can now explain CIFOL's approach to tracing things from case to case. As shown in detail in Belnap and Müller (2013a), a CIFOL sortal allows one to identify a thing falling under it, from the sortal and an extension in just one case. And typical natural sortals, such as "cat" or "lump of matter", which specify persistence conditions and a modal profile for the things falling under them, have the formal properties of CIFOL sortals.

Let us look at an example. No matter how you identify him in the case at hand (e.g., as my favourite pet, the black thing on the couch, or the best hunter on the block – all of which are case-relative descriptions), you can trace the cat, Hannibal, through all possible cases just by following the intension falling under the sortal, cat, that is identical to the extension of the identifying term in the given case. There

[9]The aim is emphatically not to provide necessary and sufficient conditions for a property to be a natural sortal property – that seems hopeless, as it is a task belonging to science and metaphysics, not to logic.

[10]Existence can be defined in terms of the special term, $*$, that signifies non-existence: $Ex \leftrightarrow x \neq *$. See footnote 8 above; for details, see Belnap and Müller (2013a).

[11]Bressan (1972) calls such predicates "quasi-absolute", the "quasi" having to do with allowing for non-existence in a case. See also the previous note.

can be just one such cat-intension (by modal separation), and the property of being a cat applies to that intension in any case (modal constancy).

Note that it is not the case that the term, "my favorite pet", falls under the sortal, cat, in the present case – that term has an intension that varies from case to case in a way that violates the persistence conditions of cats. It is true that all of my favorite pets have been and probably will be cats – but my favorite pet was first this cat and then that, and for a while I didn't have any favorite pet; furthermore, who knows whether attempts of turning me into a dog person will not be effective, or whether I won't settle for frogs in the end. "My favorite pet" does not designate a proper thing of any sort, its intension is a gerrymandered mess. But still, in the present case, the extension of "my favorite pet" is equal to the extension of "Hannibal", and the latter is a name of a cat, falling under the sortal, "cat". No matter how I identify him, given the sortal, I can trace the cat from case to case and find out about his properties in other cases.

So much for cats; the next step is to see how to link that with our discussion of possible experiments from Sect. 2. There it appeared crucial that we could trace the cat under different sortals as well – e.g., we could trace him as a massive object rather than as a biological individual. From a CIFOL point of view, we can say the following: there are other CIFOL sortals, "massiv object" and "lump of matter" among them, that also fulfill modal constancy and modal separation. In the case at hand, they do not apply to the term "Hannibal", which is after all a name for a cat: its intension falls under the sortal "cat", not "lump of matter" or such, and these sortals prescribe vastly different persistence conditions. (See our remarks in Sect. 2 about how hard it would be to trace the lump of matter that is a cat's matter in one case, from one moment to the next.) These other sortals do however apply extensionally, meaning: there is a massive object, represented by an intension falling under the CIFOL-sortal "massive object", that is, in the case at hand, identical to the cat before us – but that object is not identical to the cat in all cases. In fact, these objects come apart once the cat has taken one bite from his bowl. It is the same with the charged object identical to the cat, or with his matter: given a case, some identifying term gives us an extension, and that extension is enough, given an appropriate sortal, to identify an object of the respective sort, represented by an intension. In order to find out the cat's mass now, you have to trace the massive object now identical to the cat, to some case in which an experiment is done from which you can read off the mass; it will not do to trace the cat. Note that both objects are readily available without any fancy stories about supervenience, metaphysical overlap or contingent identity. It is not necessary to introduce any special extensions for biological individuals either – the extensions can all be taken to be perfectly physical, whatever that means. (In fact, it seems better to abstain from any verdicts about their metaphysical character, since such a verdict is not needed in order to explain their systematic place in the framework.) The answer to the question about the cat's mass is provided via a possible experiment – but that experiment needs to be done on the appropriate massive object, not on the biological individual, the cat.

5 Conclusion

We started out by following one specific use of modality in science: the link of an object's trait in a given case to a (merely possible) case in which an experiment is performed that attests to that trait. Charge or mass are generally considered to be unproblematic, categorical properties; we stuck to these in our examples so as not to become entangled in discussions about dispositions before any useful formal work could be done. We argued that the notion of tracing a thing from case to case is useful for framing the issue of possible experiments and the identity of the things involved in them.

In Sect. 3 we gave a brief overview of the standard approaches to tracing things, as represented in standard systems of quantified modal logic, and argued that they do not provide adequate elucidation of the notion of tracing. In the main Sect. 4, we presented, mostly informally, our own logical approach to the tracing of things, using the resources of the recent framework of case-intensional first order logic (Belnap and Müller 2013a). In that framework, one can specify necessary conditions for a property to be a sortal, and such sortal properties allow for tracing a thing from case to case. CIFOL's use of the method of extension and intension, together with case-relative identity, makes it possible to explain how we can trace the thing before us in a given case *both* as a cat and as a massive objects. Both means of tracing are important for some of our scientific purposes, and both find their appropriate formalization in the CIFOL framework.

Acknowledgements Research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement nr 263227, and from the Dutch Organization for Scientific Research, grant nr NWO VIDI 276-20-013. Thanks for audiences at Bertinoro, Bochum, Eindhoven and Munich for helpful discussions, and special thanks to my CIFOL co-author Nuel Belnap and to my commentator at Bertinoro, Barbara Vetter.

References

Barcan, R. 1946. A functional calculus of first order based on strict implication. *Journal of Symbolic Logic* 11: 1–16.

Belnap, N., and T. Müller. 2013a. CIFOL: Case-intensional first order logic. (I) Toward a logic of sorts. *Journal of Philosophical Logic*. Published online first; doi:10.1007/s10992-012-9267-x.

Belnap, N., and T. Müller. 2013b. BH-CIFOL: Case-intensional first order logic. (II) Branching histories. *Journal of Philosophical Logic*. Published online first; doi:10.1007/s10992-013-9292-4.

Bressan, A. 1972. *A general interpreted modal calculus*. New Haven: Yale University Press.

Carnap, R. 1947. *Meaning and necessity: A study in semantics and modal logic*. Chicago: University of Chicago Press.

Kracht, M., and O. Kutz. 2007. Logically possible worlds and counterpart semantics for modal logic. In *Philosophy of logic*, Handbook of the philosophy of science, ed. D. Jacquette, 943–995. Amsterdam: Elsevier.

Kripke, S. 1959. A completeness theorem in modal logic. *The Journal of Symbolic Logic* 24: 1–15.

Kripke, S. 1963. Semantical considerations in modal logic. *Acta Philosophica Fennica* 16: 83–94.
Kripke, S. 1980. *Naming and necessity*. Oxford: Blackwell. Originally published in 1972.
Lewis, D.K. 1986. *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D.K. 1968. Counterpart theory and quantified modal logic. *Journal of Philosophy* 65(5): 113–126.
Montague, R. 1973. The proper treatment of quantification in ordinary English. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, ed. J. Hintikka, J. Moravcsik, and P. Suppes, 221–242. Dordrecht: D. Reidel. Reprinted as chap. 8 of Montague (1974).
Montague, R. 1974. *Formal philosophy: Selected papers of Richard Montague*. New Haven: Yale University Press. Edited and with an introduction by R.H. Thomason.

The Proof Is in the Process: A Preamble for a Philosophy of Computer-Assisted Mathematics

Liesbeth De Mol

> *Mechanization tends to emphasize practice rather than theory, deeds rather than words, explicit answers rather than existence statements, definitions that are formalized rather than behavioristic, local rather than global phenomena, the limited rather than the infinite, the concrete rather than the abstract, and one could almost say, the scientific rather than the artistic.*
>
> Lehmer (1966)

1 Introduction

Is the computer really affecting mathematics in some fundamental way? Despite the historical connection between mathematics and computers, research within philosophy, history and sociology of mathematics on this question has remained relatively limited.

The main philosophical issues discussed within this context are mostly related to the challenge posed by computer-assisted mathematics to more traditional accounts within the philosophy of mathematics, accounts which view mathematics as an a priori, non-empirical and purely deductive science that generates absolute knowledge through the progressive accumulation of theorems. Computer-assisted proofs of important theorems like the four color theorem by Appel and Haken or the use of "computer experiments" to e.g. give support to important mathematical conjectures *seem* to challenge the very idea of an infallible and a priori mathematics. In this sense, studies of CaM fit in well with the growing emphasis in recent years on

L. De Mol (✉)
Centre for Logic and Philosophy of Science, University of Ghent, Blandijnberg 2, 9000 Ghent, Belgium
e-mail: Elizabeth.DeMol@UGent.be

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_2,

mathematical practices.[1] However, not all authors agree on the role of the computer here. In fact it has been argued before that, if experiments exist at all in mathematics, the computer is not (Baker 2008, p. 343) "*an essential feature of experimental mathematics. There is experimental mathematics that makes no use of computers*."

In Avigad (2008) it is argued that some of the typical questions within the philosophical literature on computer-assisted mathematics are too vague. Examples of such questions are (*id*., pp. 3–4):

- In what sense do calculations and simulations provide "evidence" for mathematical hypotheses? Is it rational to act on such evidence?
- Does formal verification yield absolute, or near absolute, certainty? Is it worth the effort?

Instead such questions should be formulated "*in such a way that it is clear what types of analytic methods can have a bearing on the answers*." The task of the philosopher is then to study how these "*pre-theoretic [questions] push us to extent the traditional philosophy of mathematics in two ways: first, to develop theories of mathematical evidence, and second to develop theories of mathematical understanding*" (*id*., p. 5). Hence, we should study such pre-theoretic questions in their proper philosophical context. Furthermore, since "*none of the core issues are specific to the use of the computer* per se [...] *issues regarding the use of computers in mathematics are best understood in a broader epistemological context*" Avigad draws two important methodological conclusions from this:

> Ask not what the use of computers in mathematics can do for philosophy; ask what philosophy can do for the use of computers in mathematics [...]
>
> What we need now is *not* a philosophy of computers in mathematics; what we need is simply a better philosophy of mathematics.

This paper nicely sums up the general tenor of some of the recent philosophical literature on computer-assisted mathematics: the object under study are issues within the philosophy of mathematics which already have a tradition and, even though the computer raises some questions that challenge more traditional accounts of philosophy of mathematics, these issues are not really essential to the use of the computer.

Even though this approach of studying CaM in a broader philosophical framework is valuable, its insistence on viewing computer-assisted mathematics as something which doesn't really change anything fundamental and merely serves existing debates, runs the risk of underestimating the actual effect on practices of CaM.

A complementary approach which does take the practice of computer-assisted mathematics more seriously seems necessary in order to get a more balanced account of the impact of the computer on (the philosophy of) mathematics. This

[1] See for example van Kerkhove and van Bendegem (2008).

has already been argued to some extent by van Kerkhove and van Bendegem (2008) where it is stated that we *should* account for the practices underpinning formal proofs, including the use of experimental methods (*id*., p. 434):

> [I]t is clear that already today mathematicians rely on computers to warrant mathematical results, and work with conjectures that are only probable to a certain degree. Every so often, we get a glimpse of what is happening back stage, but what seems to be really required is not merely the idea that the front can only work if the whole of the theatre is taken into account, but also that, in order to understand what is happening front stage, an insight and understanding of the whole is required. If not, a deus ex machina will be permanently needed.

But what does it mean to take the machine more seriously? What does it mean to give an account of "the whole theatre"? Several approaches are possible but the one I propose is one which includes a study of the technical details underpinning a practice and, as such, is bottom-up. Within this approach the computer is regarded as a real medium in the sense of people like Friedrich Kittler and Martin Carlé (Carlé and Georgaki 2013):

> The entire impact of a technically informed media theory, from matters of the vowel alphabet all the way to the realm of digital signal processing, brings about one insight: that far more than ideas, it is the 'instrumentality' of thought or the means of communication which establish a dominant regime of knowledge, thus shaping historical reality and its associate notion of truth. Media are no tools. Far more than 'things at our disposal' they constitute the interaction of thinking and perception–mainly unconsciously.

An implication of this point of view is that our mathematical knowledge is really shaped by the machine. The problems that result from its usage *must* thus be regarded as specific to the use of the computer per se. More concretely this view results in a methodology that does not shy away from the "gory" details of the (history of) computer-assisted mathematics and takes the conditions, imposed by the computer on mathematics, more seriously. On the basis of an extensive analysis of CaM, the purpose of this approach is to detect which issues *are* inherent to the use of the computer and, on their basis, to detect how the practice of mathematics is or is not affected by the computer. Such an approach is sensitive to historical fluctuations and does not aim at providing a once-and-for all given answer to the question of the impact of the computer on mathematics.

2 Human-Computer Interactions, Time-Sensitivity and Internalization

In what follows I will focus on experimental mathematics and will thus not consider issues like on-line communities of collaborating mathematicians, the impact of type-setting software on mathematics, etc.

Experimental mathematics is understood here in the sense of number theorist and computer pioneer Derrick H. Lehmer. Lehmer identifies two "schools of thought" in mathematics (Lehmer 1966, p. 745):

> The most popular school now-a-days favors the extension of existing methods of proof to more general situations. This procedure tends to weaken hypothesis rather than to strengthen conclusions. It favors the proliferation of existence theorems and is psychologically comforting in that one is less likely to run across theorems one cannot prove. Under this regime mathematics would become an expanding universe of generality and abstraction, spreading out over a multi-dimensional featureless landscape in which every stone becomes a nugget by definition. Fortunately, there is a second school of thought. This school favors *exploration* [m.i.] as a means of discovery. [B]y more or less elaborate expeditions into the dark mathematical world one sometimes glimpses outlines of what appear to be mountains and one tries to beat a new path. [N]ew methods, not old ones are needed, but are wanting. Besides the frequent lack of success, the exploration procedure has other difficulties. One of these is distraction. One can find a small world of its own under every overturned stone.

For Lehmer it is exactly this possibility of exploration that opens up the path of "*mathematics [as] an experimental science*".

In my previous research I made several detailed case-studies throughout the history of computer-assisted "experimental mathematics" to understand on a more concrete basis the impact of the computer on mathematics. These studies show very clearly the significance of technological advances in computer science (hardware, software and theoretical) for the way experiments are set-up, the types of methods that are developed and the way they are interpreted (see e.g. Bullynck and De Mol 2010; De Mol 2011): in fact, the short history of computer-assisted experimental mathematics itself already underwent important changes due to e.g. increase in computing speed, more efficient read and write operations, developments in programming etc. It are exactly these technological changes that are specific to the use of the computer and allow to trace characteristics of practices of experimental mathematics that come to the fore because of these technological conditions. These characteristics allow to partially explain the increasing popularity of so-called experimental mathematics (see Sect. 2.3).

2.1 Mathematician-Computer Interactions

If there is one characteristic inherent to the use of the computer per se, it is the interaction with the machine. Of course, there is a long history in mathematics of interactions between mathematicians and non-human instruments. The most frequently used is the pen-and-paper method: writing on a piece of paper, a blackboard etc.[2] Figure 1 illustrates the interactive feedback process of such writing practices. Evidently, such interactions are processual – one does not just have one interaction with the piece of paper but many while developing e.g. some idea for a proof or writing a result down to be communicated to the mathematical community.

[2] In this paper I do not consider earlier uses of mechanical devices within mathematics (for instance, Hartree's differential analyzer). These were much less frequently used than digital computers. A comparative study of such devices would be very interesting. It might be that some of the characteristics studied in this paper might also apply to some extent to these earlier devices.

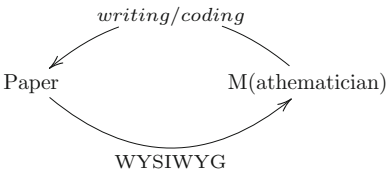

Fig. 1 Scheme of the interactive aspects of mathematical writing practices – first approximation

Within such interaction you write something down. This writing act always involves a certain level of "coding": you use symbols, drawings, abbreviations, plain text etc. This coding practice is historically determined and depends on the goal of the writing act. For instance, mathematical writings in the sixteenth century are very different from mathematical writings in the twenty-first century. Also, writing in notebooks in the process of developing for instance, a good symbolization is very different from writing a textbook. All these writing practices share the property that what you get back from the writing is usually in a WYSIWIG format: this is a term from computer science and stands for What You See is What You Get – it refers to software which uses an interface where you indeed immediately see what you get.[3] When you write or type something on a piece of paper you also immediately get what you write – a stroke is a stroke, a number is a number (at least when your pen or your typewriter are not malfunctioning).

So what happens if we transpose this scheme to interactions with a computer in the context of experimental mathematics? When you are programming a machine to tackle or explore some mathematical problem/object/process, this also involves a process of coding. However, this coding is of a very different nature when compared to "coding" on a piece of paper: whereas the "interpreter" of the mathematical writings on paper is always a human, the coding on a machine has also a non-human interpreter: the machine which executes the code. As a consequence, the coding requires a "language" that is somehow also understandable by a machine. Such intermediary language is called an interface.[4] A mathematician-computer interaction thus involves at least three components: the human, the part of the computer that processes the code and the interface. This results in the feedback scheme of Fig. 2.

This scheme produces several stages of an interaction. First of all, there is the preparatory mathematical stage which involves two substages: first, the understanding of the problem as a computational problem that can be tackled with a machine (Ia). Secondly, there is the translation of the problem into algorithms (Ib). A second stage is the translation of the algorithms to an interface – this is the actual

[3] An example of such software is Office Word where you have a direct visual of the lay-out of your text (e.g. a word in *italics* looks like a word in *italics* on your screen).

[4] Note that the use of the term "language" is a bit tricky here from a historical perspective, hence the use of double quotes. Very early machines like ENIAC did not have such intermediary language: programs had to be set up by physically cabling the machine. In this context the components of the machine, their switches and the cables used to connect them constitute the interface.

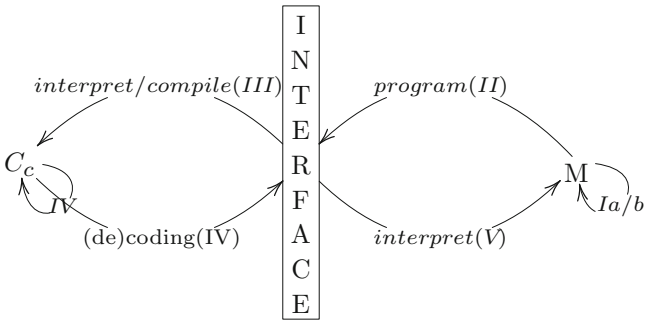

Fig. 2 Scheme of mathematician-computer interactions – first approximation

programming stage and involves the implementation of the problem. Thirdly, there is the interpretation or compilation of the program into machine code and ultimately electronic pulses (Stage III) which can then be executed resulting in an output which has a form specified by the program, e.g., a visualization, a printed (punched) table etc (Stage IV). Note that the interpretation/compilation phase in the modern sense of the word does not really apply for early digital computers. However, there are also certainly processes of translation at work in this context.[5] Finally, there is the interpretation or use of the output by the mathematician which can result in a new programming cycle (Stage V). Figure 2, however, is still a serious simplification of an actual mathematician-computer interaction. When a mathematician is using a computer to do (experimental) mathematics, the interaction is always one that develops over time and in fact involves many sequences of different interactions. For instance, there is always a process of debugging. Furthermore, it is not unusual that several different interfaces are used: the programming interface itself (some text editor like Emacs coupled with an interpreter or compiler), the interface used for the output, a debugger etc. On top of this, the communicative process between the computer and the mathematician is influenced by the communication channels themselves. This was already observed in a different context by Benoît Mandelbrot who refined Shannon's theory of communication by interpreting a communication between a sender and receiver as a game played against Nature (Mandelbrot 1953). These considerations result in more complicated feedback schemes such as the one shown in Fig. 3.

Even though this is still a simplification, this scheme indicates how complicated mathematician-machine interactions in the context of experimental mathematics actually are, involving many different stages affected by the physical and biological conditions imposed by the machine, the interface and the human (Nature). This is a very different kind of interaction when compared with that of Fig. 1.

[5]For instance, in the wiring of conditionals for ENIAC. See Bullynck and De Mol (2010) for more details.

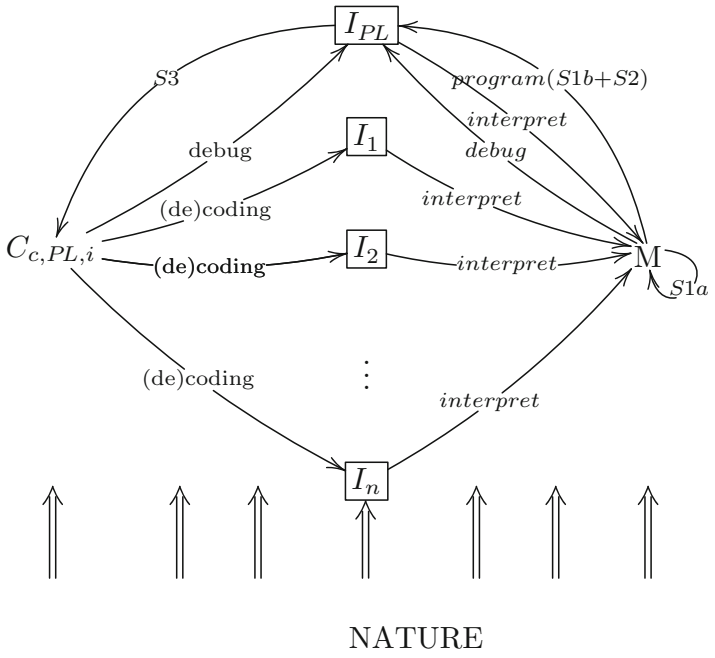

Fig. 3 Scheme of mathematician-computer interactions – second approximation

Of course, one can object against this comparison that the difference lies in the fact that the scheme of Fig. 1 does not include the actor who "reads" and, possibly, "acts upon" what is written on the piece of paper. Hence, the comparison is misleading. Indeed, the piece of paper itself is but a passive receiver of a message which can then be passed on to a second actor: the mathematician him/herself who is writing e.g. in his/her notebooks or a group of mathematicians who are reading what is written on a blackboard (without delay) or in a journal paper (with delay). However, if we were to include this second actor in Fig. 1, we would be comparing interactions between humans and non-humans with interactions amongst humans.

Such comparison between mathematician-computer interactions (MCI) and non-oral interactions between mathematicians requires that we make a detailed study of how the different aspects of MCI are affected by the fact that a non-human is involved. One telling example in this context is the stage of algorithmization: this is shaped by the fact that the algorithm is not meant for human but for machine use, the machine which is assumed to be fundamentally different from humans with its own particular "talents" to tackle a given problem.[6] Throughout the history of CaM this

[6]To put it roughly, a machine has an extremely big memory (at least nowadays) and is much more faster. Moreover it can more easily deal with certain logical complexities. A human on the other

has resulted in the development of entirely new types of algorithms (for instance, the development of pseudo-random number generators) and the transformation of existing human algorithms to machine algorithms.

One early example of this, studied in detail in Bullynck and De Mol (2010), involves the ENIAC, one of the first electronic and general-purpose machines. During a labour day weekend in 1947 Derrick H. Lehmer, his wife Emma and their children spent their time with ENIAC to compute exponents e of 2 mod p, viz. the smallest value of e such that $2^e \equiv 1 \bmod p$. It was a known fact that Fermat's little theorem could be used as a primality test. If for a given number b, $2^b \equiv 2 \bmod b$ than b is with high probability a prime number. Unfortunately, an infinite set of exceptions to this primality test exists. A table of exponents can be used to compute such exceptions. Before, Lehmer had been using Kraïtchik's tables. These tables, however, only extended to 300,000, and contained rather a lot of errors. As a result of his ENIAC computation Lehmer published a list of errors to Kraïtchik's tables and a list of factors of $2^n \pm 1$. Several different subroutines needed to be implemented on the machine including the so-called Exponent routine. Now, this exponent routine is very different from the procedure a human being would follow. A human computer would calculate the exponents e more or less in the following way. First, he would take an existing prime table to select the next prime p. Then he would calculate powers of 2 and reduce them modulo p, though not all powers, but only those that are divisors of $p - 1$. This is because a number-theorist knows that if there is an exponent e ($< p - 1$) of 2 so that $2^e \equiv 1 \bmod p$ (p prime), than e is either a divisor of or equal to $p - 1$. He might also make use of already existing tables of exponents. The ENIAC, in contrast, "*was instructed to take an 'idiot' approach*" (Lehmer 1974, p. 4). First of all, the machine needs a list of prime numbers. Of course, this list of primes could be feeded to the machine by means of punched cards, but, since this is a mechanical procedure, this would significantly slow down the computational process. Hence, it was decided not to use an existing prime table but to let the machine compute its own next value of p as it was needed.[7] The next step was to calculate the powers of 2 reduced modulo p (p being a prime) to compute exponents as follows (Lehmer 1974, pp. 4–5):

> In contrast, the ENIAC was instructed to take an 'idiot' approach, based directly on the definition of e, namely, to compute
>
> $$2^n \equiv \Gamma_n (\bmod p), n = 1, 2, \ldots.$$
>
> until the value 1 appears or until $n = 2001$, whichever happens first. Of course, the procedure was done recursively by the algorithm:
>
> $$\Gamma_1 = 2, \Gamma_{n+1} = \begin{cases} \Gamma_n + \Gamma_n & \text{if } \Gamma_n + \Gamma_n < p \\ \Gamma_n + \Gamma_n - p & \text{otherwise} \end{cases}$$

hand can rely on his/her background knowledge which he/she understands and is able to exhibit human creativity.

[7] This was done by implementing a prime sieve. See Bullynck and De Mol (2010) for more details.

> Only in the second case can Γ_{n+1} be equal to 1. Hence this delicate exponential question in finding $e(p)$ can be handled with only one addition, subtraction, and discrimination at a time cost, practically independent of p, of about 2 seconds per prime. This is less time than it takes to copy down the value of p and in those days this was sensational.

As this example shows, from the very early days of computing, it was necessary to completely transform human methods for tackling a given problem. Viz., the problem needs to be analyzed also from the machine's eye. A simple translation of the human methods would not only result in an extremely slow computation but would also require the internalization into the machine of knowledge it does not really need in order to have an efficient algorithm.

As this short analysis of MCI shows, the digital computer introduces a new type of interaction at work in mathematical practice, a practice that affects the knowledge that results from such interactions. These interactions seem to have more in common with interactions between human mathematicians than with interactions between a mathematician and his writing devices.[8] As such, we are dealing with a new social situation in mathematics which certainly is in need of further analysis.

Whereas this interactive aspect of computer-assisted experimental mathematics can perhaps be regarded as the structure through which practices of experimental mathematics are conducted, there are several other basic characteristics that are part of this interaction and affect the mathematics resulting from it. I will discuss two such characteristics here. The first is the internalization of mathematical tools and knowledge into the machine, the second is the significance of time-based reasoning within such interactions.

2.2 Internalization

If one looks at the short history of computer-assisted mathematics, it is clear that the first examples of so-called computer-assisted experimental mathematics are very different from contemporary experiments. One reason for this is that theoretical and technological advances have affected the way knowledge and skills are distributed between the human and the computer: increases in speed and memory, advances in interfaces, and advances in the "art of programming"[9] have resulted in an increasing internalization of mathematical tools and knowledge into the computer itself. There are two related aspects to such internalization: storage of information in the machine

[8] In this context, it is not surprising that mathematicians who have embraced the computer in their work have insisted on the idea of mathematician-machine collaborations. One interesting example in this context is Doron Zeilberger who has several papers with a certain Shalosh B. Ekhad as his co-author, who is Zeilberger's computer (see http://www.math.rutgers.edu/~zeilberg/pj.html).

[9] This refers to the so-called bible for programmers, viz. Donald Knuth's volumes on *The art of computer programming*.

and algorithmization. Advances in these two aspects of internalization have affected the nature of the interaction between mathematicians and computers.

During the early years of digital computing, there were two major bottlenecks. The first was the programming bottleneck, viz. the fact that it took too much time to set-up a program on a machine because there was no such thing as a programming language (De Mol et al. in press). The second problem was the memory bottleneck: the early machines did not have a large electronic memory. This implies that one could not take advantage of the electronic speed of the machine if one implemented a procedure which needed a lot of intermediary data during the computation itself. One consequence of this was the steady replacement of such tables of data by algorithms – if possible, values were not stored but computed by the machine as they were needed hence resulting in the internalization of data by means of algorithms (see the example of Sect. 2.1). If this was not possible, one had to rely on punched cards which seriously slowed down the computational process.

Another consequence of these two bottlenecks is that the interactive process mostly consisted of clearly separated phases in time. The machine was used mostly for the computational work itself, the calculation. The human took care not only of the programming but also of the exploration or consecutive analysis of the data provided by the machine. For instance, for the Lehmer computation on the ENIAC, the machine was used for the actual computation of the exponents. However, the additional work required to determine on the basis of these exponents the composite numbers was still done by a human. A similar observation can be made for the ENIAC computation of more than 2,000 digits of π and e to explore the statistical distribution of the decimal extension of these numbers. Also in this case, the main computation was done by the machine. However, the statistical analysis was done by humans. Indeed, what one typically sees with these early machines is that the interaction proceeds as follows: first, the program is prepared, then, it is programmed on and executed by the machine. This phase was often followed by one or more "debugging" phases.[10] Finally, the human does the exploration or inspection of the output which might result in a new sequence of preparation, programming, computation, (debugging) and exploration, provided there was enough machine time available. Thus, in the early years of computing, processes of internalization mostly concern the process of algorithmization either to replace the human computational work and to avoid the use of introducing large amounts of data during the computation.

This starts to change with the steady resolvement of the two bottlenecks. The availability of a bigger electronic memory together with advances in programability makes possible the steady internalization of more and more subroutines. This is for instance clear from Grace Hopper's keynote address for the ACM SIGPLAN History of programming languages conference, June 1–3, 1978 (Hopper 1981). She explains, amongst others, how important it was to develop subroutines which

[10]It is interesting to point out that for the ENIAC computation of π and e half of the programming was done to have "*absolute digital accuracy*" (Reitwiesner 1950).

were general enough to be used for a variety of purposes (e.g. a search algorithm that applies to different types) which could then be internalized into the machine (either hard-wired or programmed). One fundamental development in this context mentioned by Hopper is the significance of the machine's ability to write its own program, viz. compiling. This is a precondition for developing programming languages as intermediary languages between the human and the computer. The machine needs the ability to make its own programs in machine code when it is provided with, for instance, the following command in `LISP`, one of the oldest programming languages:

```
(*2 (/ 8 4) (+ 4 6))
```

This possibility of the machine writing its own programs was, initially, met with great skepticism (*id*., p. 9):

> Of course, at that time the Establishment promptly told us [...] that a computer could not write a program; it was totally impossible; that all that computers could do was arithmetic, and that it couldn't write programs; that it had none of the imagination and dexterity of a human being. I kept trying to explain that we were wrapping up the human being's dexterity in the program that he wrote [...] and that of course we could make a computer do these things so long as they were completely defined.

This is indicative of the transition from using computers as mere calculating aids to machines which can basically do anything an abstract Turing machine can do.[11] This relates directly to processes of internalization: the fact that the computer can do more and more and that this is being understood by humans, goes hand-in-hand with more and more subroutines being internalized into the machine. These subroutines are no longer restricted to the "pure" calculation of "raw data": they are used to visualize data, to statistically analyze data, to inspect data searching e.g. for patterns etc. In this sense, what was in the 1940s and 1950s the human's dexterity is now considered as the machine's dexterity. This development however was possible not only because of increases in programmability but also because technological advances resulted in an exponential increase in the speed by which data can be read or written in the electronic memory, both locally on individual computers as well as globally in networks of computers and humans. Indeed, it makes no sense to have a statistical tool internalized into the computer to deal with billions of data if these data themselves cannot somehow be stored internally into the machine or network of machines.

Nowadays, there exist huge libraries of subroutines not only in programming languages but also in software packages like Maple and Mathematica where "tools of dexterity" can simply be called by their name without the mathematician having

[11] From Hopper's quote one is tempted to conclude that she was not aware of developments in theoretical computer science: it was known since 1936 that there are well-defined problems that cannot be computed, provided one accepts Turing's thesis viz. that anything that is computable is also computable by a Turing machine (see Daylight (2012) for more details).

to know the complete procedures behind such names. One interesting example in this context is Sloane's on-line encyclopedia of integer sequences (OEIS) which is used by several mathematicians as an explorative tool in their work and has resulted in several mathematical papers (see http://oeis.org/wiki/Works_Citing_OEIS for over 2,000 papers that reference the encyclopedia in their work). The encyclopedia stores over 200,000 integer sequences. One very interesting feature of OEIS is that, if you have some number sequence which is not in the encyclopedia and for which you want an explanation, you can mail it to *Superseeker*. This is an algorithm which:

> tries very hard to find an explanation for a number sequence [using] an extensive library of programs that tries a great many things [...] Some programs try to find a formula or recurrence or rule that directly explains the sequence. [...] Other programs apply over 120 different transformations to the given sequence to see if any transformed sequence matches a sequence in the OEIS.[12]

This simple example indicates how processes of increased internalization affect the interaction between the mathematicians and the computer. Whereas in the early years of digital computing the division of labor was very clearly separated into the calculatory work, done by the machine, and the more "intelligent" work done by the human, this division becomes more and more blurred as more sophisticated techniques and more data are internalized into the machine. Nowadays, the machine does part of the programming, part of the inspection, etc. This changes the interaction more and more into a mathematician-machine *collaboration* or, as it has been described before by people like Licklider, a symbiosis:

> Computing machines can do readily, well, and rapidly many things that are difficult or impossible for man, and men can do readily and well, though not rapidly, many things that are difficult or impossible for computers. That suggests that a symbiotic cooperation, if successful in integrating the positive characteristics of men and computers, would be of great value.

However, the fact that more and more information and algorithms are internalized into the machine often means that they are hidden from, inaccessible to or unsurveyable for the (community of) mathematician(s). As such, this situation of increased internalization gives rise to a wide variety of new problems: how can we understand a result if part of it is hidden inside the machine? Can we trust results from the machine? Can we trust the conclusions drawn by fellow mathematicians on the basis of their experiments without having full access to the complete code and data? What does it mean to patent an algorithm? What does it mean for the community of mathematicians that they are using software packages that are not open source and which imply that it is impossible to know all the methods one is using to attain a certain result? etc. These problems lie beyond the scope of this paper but they indicate that the increased internalization of mathematical knowledge into the machine not only affects the interaction between machines and mathematicians but also results in several new problems which cannot simply be discarded.

[12]Taken from http://oeis.org/demos.html on April 5, 2013.

2.3 Time and Finite Processes

Internalization of mathematical knowledge is one aspect that results from interactions between mathematicians and computers. Another fundamental feature is the increasing significance of time and processes in mathematics.

From the early beginning onwards the fact of the speed of electronic computing was, besides its programmability, considered by many a computer pioneer as one of its greatest impacts. Hamming, a mathematician, described the effect of the significance of this electronic computing speed as follows (Hamming 1965, pp. 1–2):

> [An] argument that continually arises is that *machines can do nothing that we cannot do ourselves*, though it is admitted that they can do many things faster and more accurately. The statement is true, but also false. It is like the statement that, regarded solely as a form of transportation, modern automobiles and aerolplanes are no different than walking. [A] jet plane is around two orders of magnitude faster than unaided human transportation, while modern computers are around six orders of magnitude faster than hand computation. It is common knowledge that a change by a single order of magnitude may produce fundamentally new effects in most fields of technology; thus the change by six orders of magnitude in computing have produced many fundamentally new effects that are being simply ignored when the statement is made that computers can only do what we could do ourselves if we wished to take the time.

This speed-up in computation time is often underestimated. It is stated that the mere capability of being faster than a human doesn't change anything fundamental since, *in principle*, we can still do what the machine is doing. It is but a quantitative change. But of course, this *in principle* argument is where the catch lies as indicated by Hamming's quote: *in reality* we simply cannot do what the machine is doing. If one is really taking seriously the mathematical practice, then one *must* account for the qualitative changes that are effected by this quantitative change, else, one should neglect all such, basically, technological changes and one would end up exactly where the philosophy of mathematical practice did not want to go, viz. a largely dehistoricized mathematics which is not sensitive to external changes.

This speed-up of computations goes hand-in-hand with the fact that the objects of computers are not the traditional infinitary and stable objects of mathematics, but highly dynamic and finite processes: a computation is something that develops in and takes time. As a consequence, the computation itself can never be completely captured in the mathematical procedure to be computed and it is the task of the programmer to somehow find a way to control the dynamic processes induced by the program he/she writes: one must be able to write a program that will indeed do what we want it to do. This problem was already understood by John von Neumann who explicitly connected it to the time aspect of computations (von Neumann 1948, pp. 2–3):

> [C]ontemplate the prospect of locking twenty people for two years during which they would be steadily performing computations. And you must give them such explicit instructions at the time of incarceration that at the end of two years you could return and obtain the correct result for your lengthy problem! This dramatizes the necessity for high planning, foresight, and consideration of the logical nature of computation. *This integration of logic in the problem is a consequence of the high speed.* [m.i.]

This need for (logical) control over the behavior of the program is highly relevant in the context of computer-assisted proofs like the four-color theorem, and, more broadly, computer-assisted experimental mathematics. Usually one has thousands of lines of code which makes it extremely hard to verify that the code is doing/will do what it should do. This is the reason why it took for instance several years to review the computer-assisted proof of the sphere packing problem by Thomas Hales. It is also the reason why there is a growing need for formal proofs constructed with the help of proof-assistants like `HOL`. However, to have a formal proof one first needs a traditional proof. Furthermore, "*[i]t is a large labor-intensive undertaking to transform a traditional proof into a formal proof.*" (Hales et al. 2010, p. 3). An alternative strategy to increase the confidence in such computer-assisted results is corroboration. This was for instance proposed by Brady who proved a certain result in theoretical computer science with the help of the computer (Brady 1983, p. 662).

The fact that one needs to deal with highly dynamical processes also very often implies the irreversibility of such processes (Margenstern 2012, p. 645):

> Let us note that in our discrete time of computations, time is irreversible: it is very often extremely difficult to run an algorithm backward. At the highest level of generality it is impossible.

Such irreversibility introduces the arrow of time in the processes studied by means of the computer and, as a consequence, also in those aspects of mathematics that are studied with the help of the machine. In fact, it is this irreversibility in computational processes that has given rise to fundamental problems that resulted in new mathematical developments, for instance, the study of dynamical systems like the quadratic iterator ($f(x_i) = ax_{i_1}(1-x_{i-1})$). Such studies historically originate in the problem of error propagation during computations which became an important problem with electronic computing: since one squeezes thousands of computational steps into a feasible amount of time combined with the fact that computers are finite machines, errors resulting from truncation become highly problematic (see e.g. von Neumann 1948, pp. 3–4).

The irreversibility of computational is also reflected in the languages used to write programs. The most basic example of this is the assignment usually written as[13]:

```
var := expr
```

A simple example of this is:

```
x := x + 1
```

As explained by Margenstern (2012, p. 645), "*the notation* := *explicitly indicates that what is on the left-hand side is not the same as what is on the right-hand side,*

[13]Note that assignment is a typical feature of imperative languages. It is discouraged and sometimes even forbidden in functional languages.

and that there is a process, a consequence of which, after some time, is what we call an [assignment]." Programming languages are full of these kind of notations and, as such, introduce a notation which incorporates the processual character of computation into computer-assisted mathematics.

Such highly dynamic and often irreversible processes also mostly have the property of being unpredictable both theoretically and practically speaking. Hamming describes this unpredictability as follows (Hamming 1965, p. 2):

> One often hears the remark that *computers can only do what they are told to do*. True, but that is like saying that, insofar as mathematics is deductive, once the postulates are given all the rest is trivial. [T]he truth is that in moderately complex situations, such as the postulates of geometry or a complicated program for a computer, it is not possible on a practical level to foresee all of the consequences. Indeed, there is a known theorem that there can be no program which will analyze a general program to tell how long it will run on a machine without actually running the program.

The speed of the machine combined with the theoretical problem that one can often not predict in advance when a program will halt, if at all, implies that we cannot foresee the output. Hence, all one can do is wait and see.

This unpredictability is not just some theoretical problem or property. Indeed, it is in fact this unpredictability that usually brings mathematicians to the computer: because they cannot predict the outcome of a certain computational problem they need to rely on the machine's abilities. This often results in the need for developing local programming strategies which do not always guarantee an outcome. Since in such cases there is often the possibility of infinite programs (for instance, infinite loops) the mathematician has to make certain decisions of when a certain "program" should stop even if it is without outcome. Such decisions are informed guesses based on previous exploratory work. In all of the cases I have studied I found instances of such programs and these are often identified by the mathematicians themselves as "heuristic", "experimental" or "explorative". To give just one example, Brady, when working on his proof mentioned on p. 28 had to program the machine in such a way that it was able to differentiate between different types of infinite loops in the context of Turing machines. Such loop detection is a very difficult task since it involves infinite processes. After several computer-assisted explorations of the behavior of different Turing machines, Brady had identified two types of loops A and B and discovered a property which allowed to *tentatively* but quickly classify a given Turing machine as being a loop of type A, B or an unknown type. This was programmed as a filter called BBFILT and was described by Brady as follows (Brady 1983, p. 662):

> [It] must be remembered that the filtering was a heuristic technique based upon experimental observation.

This is also one of the reasons why he describes his "*proof techniques, embodied in programs [as] entirely heuristic*" (*id*., p. 647)

Time and processes are an inherent part of MCI: the access to the mathematical results is mediated by highly dynamical processes which introduce the problem of control over and the irreversibility of computational process, reflected in the

language used to communicate with the machine; the mathematician is confronted with the unability to predict what will come and must therefore rely on "heuristic" programming techniques and an external device to get his/her (tentative) answers. These features are part of the practice of computer-assisted experimental mathematics. They not only add an important time dimension to mathematics but even help to partially understand why mathematicians themselves often talk about "experimental mathematics" in this context.

These time-related features are part of the interaction between the mathematician and the computer and, as such, affect it. The fact that one has to wait and see during the sequences of interactions with the machine shifts this interaction further in the direction of a human-human interaction.

Dijkstra, a famous computer pioneer, in discussing the need for a formal semantics of programming languages, once explained the need for human conversation as a means to resolve semantical issues arising from human communication (Dijkstra 1961, p. 8):

> [W]e only know what we have said, when we have seen our listener reacted to it; we only know what the things we are going to say will mean in as far as we can predict his reaction. However, we only know other people up to a (low!) point and in human communication every message is therefore to a high degree a trial, a gamble to see whether the other will understand us as we had hoped. As we do not master the behavior of the other, we badly need in speaking the feed back, known as 'conversation'.

This situation also applies to a certain extent in the context of experimental computer-assisted mathematics: the humanly unpredictable processes of the computer combined with the problem of verifying that the program does what it is supposed to do also introduces uncertainty about the meaning of our own programs.[14] This is exactly why, during such interactions, we cannot simply downplay the replies by the machine as mere results of a computation which we could also have executed *in principle*. Even though we *can* have more control over the behavior of the machine than over that of our fellow human beings, we do not completely master it and as such we need its feedback not only to understand our programs but also to determine our own replies-as-programs to the machine. This is also part of the reason why machine-assisted proofs are presented in a very different manner/style than traditional proofs: since the proof results from a process of MCI in which the work of the computer is not only unsurveyable but also unpredictable and not completely controllable, the proof-as-communicated reflects this processual character of the practice that resulted in the proof. In such published proofs one indeed does not get all the details. But one does get the programs that result in it, a survey of the general structure of the proof, the strategies developed to avoid errors, etc. As such, one sees that the (communicated) proof is not some stable object but a constructive process.

[14]To be clear, Dijkstra would not have agreed on this point: according to him "*[W]e can fully master [. . .] the way in which the computer reacts*" But see in this context the quote by Hamming on p. 29.

3 Discussion

What is the impact of the computer on mathematics? From a philosophical meta-perspective the answer seems to be that nothing fundamentally changes to mathematics itself since, *in principle*, the machine can do nothing that we cannot do ourselves and it merely does what it is told to do. It is admitted that the computer does pose some new challenges for traditional philosophical problems but this merely shows that we are in need of a better philosophy of mathematics that can deal with these challenges. A serious philosophy of computer-assisted mathematics however is considered to be unnecessary.

Even though such views are perfectly arguable from a meta-perspective, they run the risk of underestimating the effects of the computer on mathematics *in reality* and on its philosophy which is itself rooted in the history of mathematics and hence sensitive to change. I am strongly convinced that it would be a missed chance for the philosophy of if it would not even make the effort of investigating more seriously practices of computer-assisted mathematics for their own sake (and less for existing philosophical debates). To this end, I have proposed an approach which takes the machine seriously as a *medium*. Such view implies that we *do* need a philosophy of the computer (in mathematics). Within such an approach, one conducts research from the bottom-up in order to trace down characteristics of computer-assisted practices which are specific to the use of the machine.

In this paper I have discussed three such characteristics: MCI, the steady process of internalization of knowledge and techniques into the machine and the significance of time and processes within computer-assisted practices. One could of course argue against this that one already has internalization of knowledge before the rise of the computer in another form, viz. by way of writing and the printed press. Similarly, one could say that since computations were already important to mathematics before the digital computer this processual nature was already part of mathematics long before the rise of the digital computer. And indeed, the claim of this paper is certainly not that there is some sudden discontinuity from what was before. What I do claim here is that these two further characteristics, as being aspects of the mathematician-computer interaction, are seriously affected by the machine and as such gain a new meaning resulting in an effect on mathematics proper. Viz., the machine has not resulted in an immediate and sudden change, but it is steadily changing features that are inherent to the practice of the mathematician, knowledge transfer, communication, collaboration, mathematical notation, etc. are changing due to the use of the computer.

Two important consequences for mathematics follow from the present discussion and show that we are in need of a better understanding of practices of CaM. Firstly, the computer introduces a new social situation into mathematics: the interaction between digital machines and mathematicians. It is striking that, on the basis of the analyses from Sects. 2.1–2.3, this interaction is shifted into the direction of communication and even collaboration between human mathematicians. Evidently, the two forms of interaction shouldn't be identified because of the active

involvement of a non-human. However, it does show that one cannot simply discard the machine as being just another tool and that further research into comparing these two modes of interaction is necessary. One obvious approach for such a comparison might be to build formal models which allow a more detailed and exact comparison.[15]

Secondly, computer-assisted mathematics explicitly and abundantly introduces time into the practice of the mathematician. John von Neumann, who was very keen on using digital machinery to study problems of applied mathematics, once stated that as mathematics (von Neumann 1947):

> travels far from its empirical source, or still more, if it is a second and third generation only indirectly inspired by ideas coming from 'reality', it is beset with very grave dangers. It becomes more and more purely aestheticizing, more and more purely *l'art pour l'art*. [...] there is a grave danger that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganized mass of details and complexities. In other words, at a great distance from its empirical source, or after much 'abstract' inbreeding, a mathematical subject is in danger of degeneration [W]henever this stage is reached, the only remedy seems to me to be the rejuvenating return to the source: the reinjection of more or less directly empirical ideas. I am convinced that this was a necessary condition to conserve the freshness and the vitality of the subject and that this will remain equally true in the future.

Far more than reinjecting empirical ideas into mathematics, perhaps the computer is reinjecting time into a discipline that has long been regarded as being above and without time.

References

Allo, P., J.-P. van Bendegem, and B. van Kerkhove. 2013. Mathematical arguments and distributed knowledge. In *The argument of mathematics*, ed. A. Aberdein and I.J. Dove, 339–360. Berlin: Springer.

Avigad, J. 2008. Computers in mathematical inquiry. In *Philosophy and the many faces of science*, ed. P. Mancosu, 302–316. Oxford.

Baker, A. 2008. Experimental mathematics. *Erkenntnis* 68: 331–344.

Brady, A.H. 1983. The determination of the value of Radó's noncomputable function σ for four-state turing machines. *Mathematics of Computation* 40: 647–665.

Bullynck, M., and L. De Mol. 2010. Setting-up early computer programs: D. H. Lehmer's ENIAC computation. *Archive for Mathematical Logic* 49: 123–146.

Carlé, M., and A. Goergaki. 2013. Re-configuring ancient Greek music theory through technology: An adaptive electronic tuning system on a reconstructed ancient Greek barbiton. In *La musique et ses instruments* [Music and its instruments], ed. M. Castellengo and H. Genevois, 331–380. Paris: Editions Delatour.

Daylight, E. 2012. *The dawn of software engineering: From turing to Dijkstra*. Heverlee/Belgium: Lonely Scholar.

[15]See Allo et al. (2013) where epistemic logics are used to study processes of finding proofs by communities of mathematicians.

De Mol, L. 2011. Looking for busy beavers. A socio-philosophical study of a computer-assisted proof. In *Foundations of the formal sciences VII. Bringing together philosophy and sociology of science*, Studies in logic, vol. 32, ed. K. Francois, B. Löwe, Th. Müller, and B. van Kerkhove, 61–90. London: College publications.

De Mol, L., M. Carlé, and M. Bullynck. in press. Haskell before Haskell: An alternative lesson in practical logics of the ENIAC. *Journal of Logic and Computation*. doi:10.1093/logcom/exs072.

Dijkstra, E.W. 1961. On the design of machine independent programming languages. Report MR34, Stichting Mathematisch Centrum, Amsterdam.

Hales, Th., J. Harrison, S. McLaughlin, T. Nipkow, S. Obua, and R. Zumkeller. 2010. A revision of the proof of the Kepler conjecture. *Discrete and Computational Geometry* 44: 1–34.

Hamming, R.W. 1965. Impact of computers. *The American Mathematical Monthly* 72: 1–7.

Hopper, F.M. 1981. Keynote address. In *History of programming languages*, ed. R.L. Wexelblat, 7–20. New York: Academic.

Lehmer, D.H. 1966. Mechanized mathematics. *Bulletin of the American Mathematical Society* 72: 739–750.

Lehmer, D.H. 1974. The influence of computing on mathematical research and education. In *Proceedings of symposia in applied mathematics*, vol. 20, ed. J. Lasalle, 3–12. Providence: American Mathematical Society.

Mandelbrot, B. 1953. *Contribution à la théorie mathématiques des jeux de communication*. Paris: Laboratoires d'électroniques et de physique appliquées.

Margenstern, M. 2012. Comment. In *A computable universe*, ed. H. Zenil, 645–646. Singapore: Worldscientific.

Reitwiesner, G.W. 1950. An ENIAC determination of π and e to more than 2000 decimal places. *Mathematical Tables and Other Aids to Computation* 4: 11–15.

Van Kerkhove, Bart, and Jean Paul van Bendegem. 2008. Pi on earth, or mathematics in the real world. *Erkenntnis* 68(3): 421–435.

von Neumann, J. 1947. The mathematician. In *The works of the mind*, ed. R.B. Heywood, 180–196. Chicago: University of Chicago Press.

von Neumann, J. 1948. Electronic methods of computation. *Bulletin of the American Academy of Arts and Sciences* 1: 2–4.

The Future Role of Computation in Science and Society

Patrick Suppes

1 Some Examples of Large-Scale Computation

Let me begin with three simple examples, well, perhaps they are not so simple. The first consists of the extreme demands for large-scale computation of data coming from the Large-Scale Hadron Collider (LHC) at the laboratory of Conseil Europeen pour la Recherche Nucleaire (CERN).

The LHC produces at design parameters over 600 millions collisions ($\sim 10^9$) proton-proton collisions per second in ATLAS or CMS detectors. The amount of data collected for each event is around 1 MB (1 Megabyte).

$$\begin{aligned} 10^9 \text{ collisions/s} \times 1 \text{ Mbyte/collision} &= 10^{15} \text{ bytes/s} \\ &= 1 \text{ PB/s (1 Petabyte/second)} \end{aligned}$$

Since 1 DVD ~5GB: 200,000 DVDs per second would be filled, or about 6,000 IPods (ones with 160 GB of storage) per second!

A trigger is designed to reject the uninteresting events and keep the interesting ones. For example, the ATLAS trigger system is designed to collect about 200 events per second.

$$200 \text{ events/s} \times 1 \text{ Mbyte} = 200 \text{ MB/s}$$

Taking two shifts of 10 h per day, and about 300 days per year:

$$200 \text{ MB/s} \times 2 \times 10 \times 3,600 \times 300 \sim 4 \cdot 10^{15}$$

$$\text{bytes/year} = 4 \text{ PB/year}$$

P. Suppes (✉)
Center for the Study of Language and Information, Stanford University, Ventura Hall, Stanford, CA 94305-4115, USA
e-mail: psuppes@stanford.edu

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_3,

Collectively, the LHC experiments produce about 15 petabytes of raw data each year that must be stored, processed, and analyzed.

The second example is the astronomers' square kilometer array (SKA). Astronomers will need a top ranking supercomputer to combat the data deluge from SKA. The amount of computer data generated by the entire world in 2012 will need to be stored in a single day for the world's most powerful telescope – the Square Kilometre Array (SKA) – and the International Centre for Radio Astronomy Research (ICRAR) is gearing up to meet that unprecedented need.

ICRAR scientists say the $2 billion SKA will generate one exabyte of data – a million terabytes (or one quintillion bytes) – every day while it searches the sky with the power to detect airport radars in other solar systems 50 light years away.

There is a potentially important use of the computing power of SKA. This kind of computational focus is essential to make accurate predictions of the paths of meteorites or other bodies entering the solar system, some with a possible trajectory close to that of Earth. The better the forecast of the past of the body headed toward the general trajectory of Earth, the better the chance of avoiding a disaster of major proportions.

On February 15, 2013 there was a massive splintering of a meteorite over Siberia which would have caused enormous damage if it had been over a major metropolitan center. This kind of future event would focus the entire population of our planet, if it were to have a magnitude considerably greater than that of the recent Siberian meteorite. It is somewhat surprising that until recently astronomers paid little attention to this kind of possibility. But computation limitations played a role in the neglect.

It was understood how difficult it is to make observations sufficiently early to forecast with any real accuracy the trajectory of a meteorite, comet or other foreign body approaching Earth. With SKA, we now have that ability more than at any other time in the past. No doubt, in the future better efforts will concentrate even more on this kind of problem. Another aspect of this example of the need for large-scale computation is the sheer number of potential candidates for the cause of such a disastrous event.

This example exhibits well a feature of what we may expect from ever better methods of computation. On the one hand, we will learn much more about the world we live in and understand it better. Many positive things will follow as a result. But we will also have a deeper and a more troubling view of the disasters that may lie ahead in the not-too-distant future.

Here is a third example from meteorology.

In 1963 Edward N. Lorenz (a meteorologist and a mathematician) wrote a remarkable paper. Lorenz's search for a three-dimensional set of ordinary differential equations which would model some of the unpredictable behavior which we normally associate with the weather. The equations, which he eventually hit upon, came from a model of fluid convection. They are

$$\frac{dx}{dt} = \sigma\,(y - x)$$

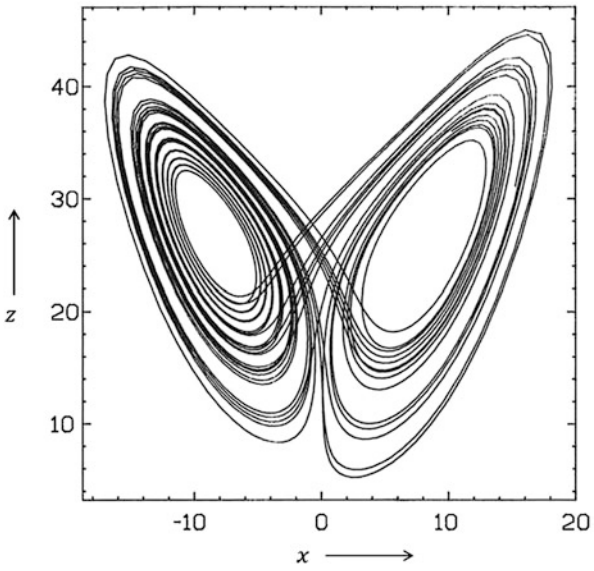

Fig. 1 A numerically computed solution to the Lorenz equations projected onto the x, z plane $\sigma = 10$, $b = \frac{8}{3}$, r = 28.0

$$\frac{dy}{dt} = rx - y - xz \qquad (*)$$

$$\frac{dz}{dt} = xy - bz$$

Where σ, r and b are three real positive parameters.

Briefly, the original derivation (Lorenz 1963, 1979) can be described as follows. A two-dimensional fluid cell is warmed from below and cooled from above and the resulting convective motion is modelled by a partial differential equation. The variables in this partial differential equation are expanded into an infinite number of modes, all but three of which are then set identically to zero. The three remaining modes give the equations (*). Roughly speaking, the variable x measures the rate of convective overturning, the variable y measures the horizontal temperature variation, and the variable z measures the vertical temperature variation. The three parameters σ, r and b are respectively proportional to the Prandtl number, the Rayleigh number, and some physical proportions of the region under consideration; consequently, all three are taken to be positive.

For wide ranges of values of the parameters, approximate solutions to the equations (*), calculated on a computer, look extremely complicated.

Figure 1 shows the projection onto the x, z plane (y = constant) of one such solution, when $\sigma = 10$, $b = \frac{8}{3}$ and r = 28. Note that the trajectory shown does not intersect itself if we consider the full three-dimensional picture. The crossings in Fig. 1 are the result of projection onto two dimensions.

These first examples are exotic. They show the new directions in science at their most extreme. In fact, the Lorenz-type example is meant to be a negative one, which shows that in principle there is no hope for predicting the weather in detail at any long range. This conclusion is now pretty generally accepted for our present regime of scientific meteorology without a scientific revolution that we cannot now foresee. It is reasonable to say that it has been shown decisively that the limitations of science are reflected in everyday life in no better way than they are in limitations on prediction of the weather, for even 2 or 3 weeks in advance. The source of this negative finding is easy to locate. In a general conceptual way, it is the enormous complexity of the weather system itself. It is not some simple Newtonian system of a few particles, say two, the most which we can study in the greatest possible detail. (It is important to remember that the general theory of even three Newtonian particles is unmanageable.)

It is hopeless to predict the behavior, with any precision, of any actual weather system for an extended period of time. The outcome of this clear example is the conclusion that we are optimistic, indeed astonishingly overly optimistic, if we think that all the major problems of modern science can, without doubt, in due time be solved by sufficiently powerful computation methods.

This extended example of a simplified model of the convection of the atmosphere illustrates an important strategy in using complicated computation models to prove that a simplified model of something like the atmosphere cannot be fully understood by the methods of analysis mathematically currently available. If such simplified examples can be shown to be impossible to find a solution for, then it surely follows that the full system will exhibit a similar impossibility. The virtue of such simplified models is that we can study very thoroughly the unpredictable behavior. Usually the sources are parameters enormously sensitive to initial or boundary conditions, but not entirely. Turbulence, which contributes to the unpredictable behavior of the weather, is an example that in its unpredictability is not entirely dependent on initial and boundary conditions, but on the motion of the fluid itself.

What such negative examples as those of Lorenz show is that the over optimism of popular predictions about the continued success of science are often exaggerated. Not enough attention is paid to how difficult it is to solve any real problem of any complexity. My favorite example is the Newtonian mechanics of point particles. A simpler physical model of any interest can scarcely be thought of. But, as mentioned earlier, consider the situation as it is today, we have a superb understanding of the behavior of one particle by itself with specified forces, and a good understanding of two particles forming an isolated system with Newtonian forces of mutual interaction. But even this isolated kind of system is in general unmanageable, from a computational standpoint as soon as we reach the problem of predicting the behavior of three such particles. In terms of the ordinary affairs of the world, this seems absurd, and yet this is an important and now well recognized limitation of that stronghold of deterministic science, classical particle mechanics. No doubt in the decades ahead, we shall learn more and more about the behavior of such systems, and yet the results of Lorenz and others show that, within the

present framework of mathematical thought, a full and complete solution will not be possible for most problems one might formulate. The problem of predicting accurately the trajectories of foreign bodies coming close to Earth is a vivid example of the unexpected and strong limitations of the predictability of even what are conceptually very simple systems. The point of mentioning such systems here is to make clear that even the most massive large-scale computational methods we can now envisage will not solve many simple problems in any complete form.

Let me conclude this discussion of scientific examples with some more positive ones, for which we expect to make considerable progress in the reasonably near future. The purpose of these examples is to give a balanced picture of the future of computation. All of these examples, as well as the earlier ones, depend upon computations but in these last set of instances I predict the future is positive, because the demands for computation are not too difficult to meet.

Perhaps the easiest way to find a large number of positive examples is to look at the revolution that is occurring in all parts of medicine and associated health sciences. One computational aspect of direct importance in the modern digital world is that the sample sizes of testing new drugs can be managed and increased by several orders of magnitude. At relatively little cost we have just begun to learn how to use digital data on a given medical problem by looking at many millions of medical records in a short period of time. What can be done now would have been unthinkable even 20 years ago. In fact, one of the most noticeable features of this work is its widespread international character. We can read daily about diseases spreading in Africa, China, Russia, Norway or any other part of the world about as easily as we can read about diseases among our current neighbors.

A second matter that is less in the press but of great importance is the improvement in surgical procedures, many of which are based on careful computations, or often on the interpretive use of digitally based imaging devices which, either for their construction or their use themselves, depend upon massive computations.

A third medical example is the small computer chip of a pacemaker that can be inserted in 15 min of surgery and that can play such a radical role in the cardiovascular health of the patient. Again we are dealing with technology that would have surprised everyone 40 or 50 years ago in its power and simplicity of implementation. From all indications, this is only a simple example of the complicated process underway of introducing evermore artificial computing power into our bodies. With nanotechnology, we can now think in a practical way of computers in the bloodstream monitoring closely and continually many subtle chemical and physical properties, which information can be easily observed and automatically transmitted to the appropriate medical workstation.

Within the range of computations that have been discussed, perhaps the most exciting and significant development is a current well-supported effort by governments and many private foundations to understand how the brain computes. Because of my own interests, I will try to say something more detailed about this area of research.

1. The system signaling is electromagnetic.
2. A reasonable hypothesis is that collections of neurons synchronize to approximate weakly coupled electromagnetic oscillators to do system computation and signaling.
3. I focus on one major problem: the physical account of brain computations in talking, listening, writing and reading, particularly on verbal memory storage and retrieved.
4. There are literally thousands of psychological papers on this problem, but no detailed physical models.
5. This is what I call abstract psychology.
6. Here is a sketch of one physical model on using phase to recognize English phonemes in the brain.
7. EEG experiment with thousands of trials and about 32 GB of data.
8. Signals seem composed of waves between 2 and 9 Hz.
9. Each electromagnetic sine wave has a frequency, phase and amplitude.
10. Amplitude is of little use. So phases of frequencies have a pattern for a given phoneme such as *p, b, t*, etc.
11. Here are the mean phase brain patterns for four frequencies of six initial phonemes (Wang et al. 2012).
12. This is only a beginning, but promising.
13. There is also progress on the semantic side in terms of semantic associative networks, but no time for any details here.

I perhaps have not stressed enough how very far we are from thoroughly understanding the computations of the brain that are essential to our daily activities. What is needed most, and because of its importance I emphasize it again, is an understanding of the physical mechanisms that do the continual computations required for walking, talking, listening, and in general perceiving what is going on around us, not as a static picture, but as a continually changing environment affecting not only what we see, but what we hear and touch as well (Fig. 2).

No doubt, given the complexity of the system, we will soon be seeing theorems on the impossibility of having a complete theory of the computations. As in the case of the weather, but of a still more urgent nature, we will press on, determined to find approximations that are ever more refined as research continues. But at no time in the near future are we going to have anything like a complete understanding of how we are producing or listening to the stream of natural language in ordinary conversation at about three words per second. At first glance, the computational processes required to support the easy and natural activities of talking and listening seem out of reach. But just as in the case of the weather, without being foolish in our predictions and too pessimistic about our findings, we will discover much that is fundamental and that can be of great use in perfecting, if nothing else, the conversations between us and our devices.

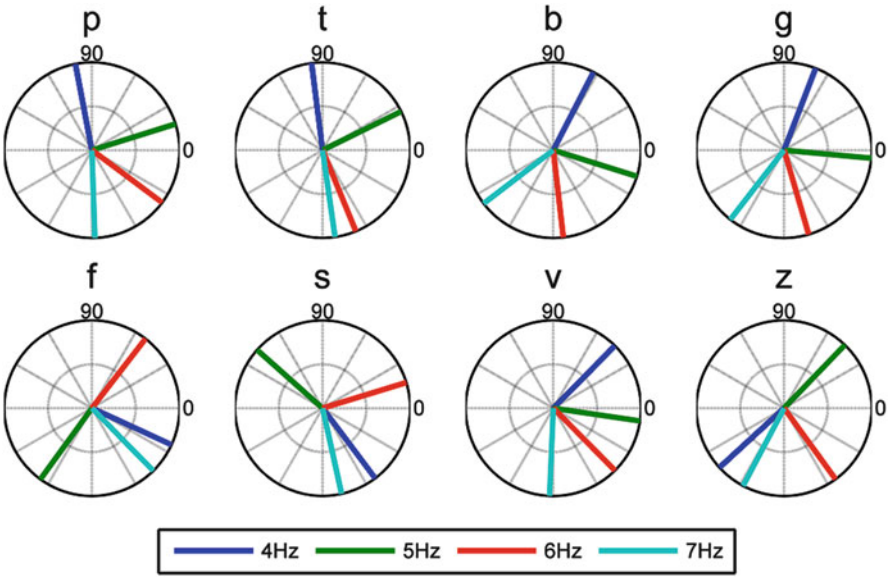

Fig. 2 Brain phase pattern of six phonemes

2 From Science to Society

Here I discuss two problems that may not be at the very center of scientific research, but are of the greatest importance to society. The first is the possibility of the indefinite extension of human life, and the second is the end of work for most people.

Let us begin with some data on the world's oldest humans. The record setter, Jeanne Louise Calment, at death was 122 years old. She was born February 21, 1875 and died August 4, 1997.

Eight of the last nine world's oldest persons at death were all under 120 years. The most prominent group were 6 at 114 years.

Let us look at Japan, a keeper of excellent records. In 1990, there were 3,000 Japanese more than 100 years old. The oldest was 114. In 2010, 44,000 Japanese were more than 100 years old, and again the oldest were 114.

Some different data come from experiments and studies on calorie reduction (CR), which was first studied in 1934. CR alone led to at least 50 % life extension of short-lived animals, such as mice (in 1934) and later many kinds of insects. Of almost as great importance; the animals and humans had better health as well.

In 1989 University of Wisconsin scientists started a study of 20 adult male rhesus monkeys with 9-CR, 11-normal subjects. In 2009 it was too early for life extension results, but the CR monkeys were much healthier.

These two kinds of arguments plus others that are easily given, but I have omitted, lead to the following conjectures or proposals for extending human life.

1. Use of stem cells to replace failing organs.
2. Careful monitoring of input of various chemicals, such as calcium.
3. More fundamental changes in gene and protein structures.
4. The big question: Replacing old brains with new ones.
5. Totally exaggerated hope: full body replacement.

I now turn to new technology and the end of jobs.

> The role of humans as the most important factor of production is bound to diminish in the same way that the role of horses in agricultural production was first diminished and then eliminated by the introduction of tractors.
>
> Wassily Leontief and Faye Duchin,
> *The Future Impact of Automation on Workers, 1986*

I support the view of this paragraph with a variety of considerations and data.

1. The technology revolution is replacing human beings with computers, robots and other intelligent devices in every sector of employment in the global economy. As is well-known, millions of workers have already been eliminated from many kinds of labor, ranging from agricultural workers to bank clerks. The handwriting is on the wall, as computers improve and become more intelligent, the number of jobs for which humans are needed will be negligible compared to the numbers employed at present.
2. Here is an amusing, but significant kind of example. In order to avoid various kinds of terrorism of human workers, Israelis have developed a melon picker (ROMPER). It uses special sensors to determine whether a melon or other type of fruit is ripe to pick.

   Similar robots have been developed, and are getting smarter by the day, to plow and seed fields, feed dairy cows, pigs and other stock. It is predicted that by the second half of this century we will see a wide-spread development of fully automated and computer-driven factory farms.

   Note that the decline of the number of agricultural workers in the United States and other advanced economies, in the twentieth century, was the most significant change in agriculture yet seen anywhere in the world.
3. No more factory workers. In the last part of the twentieth century, Japan was famous for its nine auto makers producing more than 12 million cars per year with fewer than 600,000 workers. In contrast, Detroit auto makers employed more than 2.5 million workers to produce about the same number.

   It is a common prediction that, by the end of this century, there will be no workers in the automobile factories, but only a small number managing the ever larger number of robots. Moreover, it has been widely noted that these robots work 24/7, without benefits, coffee breaks, medical illnesses or demands for pay increases.
4. By the end of this century, blue-collar workers will be gone in the developed countries, and fading away in the rest of the world.

5. With modification, of course, these same remarks apply to the service industries, banking, insurance and all kinds of retail and clerical work.

Here is my last topic, back to the past: A future without work with a population only of Aristocrats having a future without work.

In Jane Austen's, *Pride and Prejudice* (1959), to mark her social status Elizabeth Bennet sharply responded to Mr. Darcy that her father was just as much a gentleman as he. Will we return to the life of aristocrats in the eighteenth and nineteenth centuries, of ladies and gentlemen, who would not dirty their hands in commerce, as the phrase went, or in any other kind of productive labor?

Will humans become a breed of amateur managers telling their smart robots and other devices what to do, but only in the most general terms that reflect ignorance of the technical details that must be mastered by these new "devices", not by human workers? If this is what really comes about, then the nature of the change will indeed be dramatic. The new masters of Earth will be smart robots telling all of their devices and humans, as well, what to do, or rather, as in the case of the Aristocrats of Jane Austin's days, what will be accepted as proper behavior of a non-working human Aristocrat.

At the same time, these smart robots will be controlling their own revolution and moving ever further away from satisfying any natural concept of humanity. In this process, humans will become as extinct as dinosaurs are now.

I am sorry to say this seems the most likely future, but there remains a small probability that we, as humans, can stay ahead of the games and be Aristocrats in the intellectual style of Euclid, Apollonius, Claudius Ptolemy, the unknown Mayan astronomer, Sir Isaac Newton, Henry Cavendish, Pierre-Simon Laplace, James Clerk Maxwell, and the many others who belong on this distinguished list.

Like competitors today from China to California and Cambridge, the smart robots of tomorrow will engage in a dynamic competition to see which can evolve and develop the deepest new scientific concepts. My bet is still on the odds favoring the robots. The only saving thought is that what might still be the actual outcome will be a wonderful hybrid species of thinking creatures, half human and half robotic, with powerful computation potential on board, and truly amazing further remote computational power easily and quickly available.

To end, here are some new philosophical questions generated by this discussion of computation in the future.

1. Should we accept an economically supported life of leisure for all those who do not want to work?
2. Are there serious normative negative arguments against life extension if possible?
3. Should we normatively argue for the desirability of less people, given life extension and end of routine work?
4. But in stronger terms, is there a meaningful optimum size of population, as work for humans decreases and almost ceases, yet life extension continues?
5. Do our fundamental concerns for freedom of action and thought, as well as the rule of law, apply with some possible changes, to robots as well?

References

Austen, J. 1959. *Pride and prejudice*. New York: Dell.

Leontief, W.W., and F. Duchin. 1986. *The future impact of automation on workers*. New York: Oxford University Press.

Lorenz, E.N. 1963. The predictability of hydrodynamic flow. *Transactions of the New York Academy of Sciences Serial II* 25(4): 409–432.

Lorenz, E.N. 1979. On the prevalence of aperiodicity in simple systems. In *Global analysis*, ed. M. Grmela and J.E. Marsden, 53–75. New York: Springer.

Wang, R., M. Perreau-Guimaraes, C. Carvalhaes, and P. Suppes. 2012. Using phase to recognize English phonemes and their distinctive features in the brain. *Proceedings of the National Academy of Sciences* 94(5): 20685–20690.

In No Categorical Terms: A Sketch for an Alternative Route to a Humean Interpretation of Laws

Kerry McKenzie

1 Introduction

A debate over the modal status of natural laws and physical properties has been raging in the literature for decades, and the factions are by now familiar. On the one hand we have the *Humeans*, primarily represented by Lewis and his followers, who commit to the idea that the fundamental properties are categorical in character and the laws of nature metaphysically contingent (see, e.g., Lewis 1983; Loewer 1996; Cohen and Callender 2009). On the other hand we have the *Anti-Humeans*, such as Ellis and Bird, who believe that the fundamental properties are essentially dispositional and the laws metaphysically necessary (see Ellis 2001; Bird 2007). As the titles suggest, what ultimately divides the two factions is their attitudes to Hume's famous dictum that there are no "necessities in nature" – that is, no necessary connections between distinct existences that cannot be reduced to those of mathematics or logic.

This binary opposition in modal disputations will be familiar to many of us (as a first approximation at least). For brevity, I will designate the debate between these two sides, so characterized, as the "canonical debate". Despite the familiarity and prominence of this debate, however, I want to argue that there are profound problems afoot in the presuppositions made by each side, and in particular that the basic terms in which it is conducted are woefully out of date. Both parties, after all, have pretensions to giving an account of *fundamental* laws and properties, and both of these are obviously the sort of thing that will be described by fundamental *physics*,

K. McKenzie (✉)
Department of Philosophy, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Alberta, Canada
e-mail: mckenzie_kerry@ymail.com

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_4,

assuming that they can be described anywhere at all.[1] But the impression given by the major works of this debate is that Coulomb's law represents the cutting-edge of modern physics, which it emphatically does not. Nor is this a merely technical and metaphysically uninteresting point, made by another philosopher of physics lamenting the lack of engagement by contemporary metaphysicians with contemporary physics; for on the contrary, as we shall see, incorporating concepts basic to current fundamental physics can change the modal landscape quite dramatically, and naturalism demands of us that the structure of our metaphysical debates must change along with it.

In the course of arguing for that conclusion, my focus will be kept on the fundamental *kind* properties throughout – that is, the state-independent properties such as charge, hypercharge, isospin, etc. that we take to define the various fundamental particle kinds.[2] The critical part of my argument will proceed in two stages:

> firstly, I will argue that the canonical debate over nomological modality assumes an account of natural law that is not appropriate for elucidating either fundamental properties or fundamental laws, and
>
> secondly, if we move to a more realistic account of fundamental laws, then it is no longer clear that there is any place for categorical properties in our metaphysics, or at least not as such properties are standardly conceived; nor that one may regard the fundamental laws as contingent in anything but a tightly circumscribed sense.[3]

Since the canonical Humean package is defined by a commitment to the categorical nature of fundamental properties and the contingency of natural laws, the tempting conclusion to draw from all this is that Humeanism is dead in the water. Surprising as it may at first seem, however, I will argue that such a conclusion would be too rash. In particular, I will argue that modern physics makes space for a view in which necessitarianism about laws can be combined with a broadly Humean outlook; it is a Humeanism that must jettison its commitment to categoricalism, granted, but I think that we can claim it as a viable Humeanism about laws nonetheless.

To begin, then, I will outline in a little more detail what I take the canonical debate over laws, properties and modality to consist in. Once that is in place, I will articulate the problems I perceive to be inherent in the act of appealing,

---

[1] That the debate is principally over *fundamental* properties is emphasized in Bird op. cit.; see, e.g., Sect. 3.3 and the discussion of "finkish dispositions".

[2] The property of mass requires a separate treatment, as my discussion will be limited to so-called *internal* symmetries throughout; it is essentially because we do not currently possess a working quantum-theoretic treatment of gravity that accounts for this exceptional status. A longer paper would expand upon the relevant distinctions in a fuller way than I can here.

[3] In particular, I will argue that there is reason to think that the variation permitted is limited to the values of numerical constants only.

in the fundamental physics context, to the notion of categorical properties, and also in maintaining that the fundamental laws can be regarded as metaphysically contingent.

2 The Canonical Account of Laws, Properties and Modality

Painting things in as broad brushstrokes as possible, there are two categories of modal accounts of laws to be found in the literature. On the one hand we have *contingentist* accounts, according which the laws of nature consist of metaphysically contingent connections between properties. The majority of contemporary adherents of this view adopt some version of Lewis' "sophisticated regularity" view. On the other hand we have *necessitarian* accounts, according to which the connections between properties are regarded as metaphysically necessary – or at least, metaphysically necessary in the sense that properties obey identical laws across any and all possible worlds in which they are instantiated.[4] This dispute over whether natural laws are contingent or necessary is the bone of contention regarding laws in the canonical debate.

The next thing to note about the canonical debate is that each of these modal accounts of laws is grounded in a prior modal conception of properties.[5] Thus, on the one hand, necessitarians about laws typically assume an account of fundamental properties according to which they are "essentially dispositional". Since part of what it is to be an essentially dispositional property is to imply instances of laws, it follows on this view that a given species of fundamental particle, defined by a given set of fundamental kind properties, can act in accordance with *one and only one* law across all possible worlds in which tokens of it are instantiated. It is thus this modal conception of properties that necessesitarians typically take to *account* for the fact that the laws are metaphysically necessary. On the other hand, contingentists about laws reject this view of fundamental properties, and as such also reject the idea that the kinds instantiating such properties bring in their wake a unique law. They rather endorse an opposing view in which fundamental properties are deemed "categorical", and it is this categorical conception of properties that underwrites the idea that a given kind of particle could behave differently.

In what I call the "canonical account", then, contingentism about laws goes hand in hand with categoricalism about properties, and necessitarianism about laws with dispositional essentialism. The package consisting of the first pair of commitments defines the *Humean* perspective on the fundamental laws and properties in the canonical account; the package consisting of the second pair defines the

[4] Such a conditioned necessity is sometimes called "weak" metaphysical necessity.

[5] As Earman and Roberts (2005) point out, this has the result that both factions are committed to a thesis in which the laws supervene on the fundamental property basis, and as such don't strictly speaking "do any work". See also Mumford (2004).

Anti-Humean point of view.[6] Since it is in each case a modal commitment regarding properties that grounds the corresponding modal stance on laws, the modal characterization of properties is really the fulcrum about which the canonical debate turns. This circumstance is reflected in the fact that so much of the ink spilled by Anti-Humeans concerns the proper articulation of essentially dispositional properties.

However, and while what exactly is involved in the concept of an essentially dispositional property has been discussed at great length, I think we have to agree with Mumford when he says that "it is quite difficult to find, anywhere in the literature, a specification of what exactly is intended by 'categorical property"' (Mumford 1998, p. 75). But it is obvious that without *some* such specification, the precise connection between categorical properties and the contingentist interpretation of laws can only remain murky, and with it the content of the Humean package as a whole. If one looks at the literature, what one finds in lieu of a precise characterization is that a variety of strategies are deployed to at least gesture at what is intended here. One finds categorical properties characterized, for example,

> in *metaphorical* terms, as those that don't 'look outward to interactions', or as those properties that don't 'point beyond' themselves; those that are 'self-contained . . . keeping themselves to themselves' (Armstrong 1993, pp. 69, 80); or alternatively
>
> in explicitly *nomological* terms, as those properties that are 'free of nomic commitments' (Carroll 1994, p. 8), hence as those that do not 'necessarily involve laws' (Loewer op. cit., p. 200); or sometimes
>
> in *spatiotemporal* terms, namely as those properties such that 'their instantiation has no metaphysical implications concerning the instantiation of fundamental properties elsewhere and elsewhen' (Loewer op. cit., p. 177).[7]

---

[6] A couple of points are due regarding my designation of the above as the "canonical debate". (1) The first point is that, of course, not every protagonist in the literature on natural modality fits neatly into one or other of the categories just defined. Armstrong's view, for example, commits to both categoricalism about properties and contingentism about laws, but nonetheless endorses primitive necessary connections (albeit "soft" or "contingent" ones) between the properties involved in laws, apparently in conflict with Hume's dictum (Armstrong 1983). As such, his position is often taken to lie somewhere between the two above positions, and designated as "semi-Humean" (see for example Bird op. cit., Sect. 1.1). However, given the shared commitments between Armstrong and the Humean position as defined above, my criticisms of the latter position will apply to Armstrong's account too. (The relevance for Armstrong's concerns of my conclusion regarding the vindication of Hume's dictum is something I cannot discuss here.) (2) Another prominent party in this debate that may not seem to fit neatly into above-defined categories is Mellor. Mellor is inclined towards a contingentist view of laws (see e.g. his 2000, p. 770 – but note the qualification on p. 772), and yet is ambivalent as to whether properties should be thought of as categorical (see Mellor 2003, p. 231). It therefore appears that one can be a Humean about laws without committing to categorical properties. However, Mellor's ambivalence about categorical properties seems to owe largely to the lack of clear and consistently-used criteria for what "categorical" means in the first place, and I will have something to say about this in the next section.

[7] Another way to characterize categorical properties is in terms of quiddities: properties will be said the be categorical if their identity is exhausted by their quiddity. However, since it is this lack of any other features that implies each of the above designations, problems for any of the above characterizations will also be problems for this one.

There thus seem to be a number of ways of approaching what is meant by "categorical property". Greater variety does not equate with greater clarity, however, and it would certainly be nice if what is meant by "categorical" in this context could be sharpened up. Let us therefore try to do so now.

A strategy frequently adopted to convey in concrete terms that which is meant by "categorical" is that of simply *conveying by example* the implications that such properties have for some familiar natural laws. So, for instance, it is often cited that on the categoricalist view charged particles are not bound to obey Coulomb's law, and in particular that "negative charges might have been disposed to repel positive charges, or some other relation may have held between them" (Bird op. cit., p. 68). Thus part of what is meant by calling charge "categorical" is that

$$F(x, y) = +C\frac{q(x)q(y)}{r^2(x, y)}$$

– Coulomb's law with a sign flip – represents a possible law. Similarly, it has been said that if charge is categorical then "the contribution of distance might have been such that an inverse cube law held" instead of the Coulombic inverse square, so that

$$F(x, y) = -C\frac{q(x)q(y)}{r^3(x, y)}$$

is also taken to represent a possible law on this view (Armstrong 2005, p. 313).[8] Now, the first thing to point out about this strategy is that the specific examples offered in the literature of alternative laws are typically rather conservative in how they differ from the actual laws – consisting in these cases just of a sign flip and unit increase of power in the denominator respectively. The second thing to point out is that such discussions tend to be utterly silent on what *principles govern* how, and how considerably, the actual laws may be tinkered with so as to generate acceptable other-worldly alternatives such as these, and thus also to be silent on the range of nomic possibilities that are open to a given property. Insofar as the concept of categorical properties is to be articulated in terms of this variation, then, it is obvious that it will only remain unclear pending some statement of what principles limit how a law can be tinkered with so as to generate possible alternatives. However, given that such properties are explicitly regarded as "free of nomic commitments", perhaps we should take this absence to indicate that it simply goes without saying that there *are* no such principles (or at least no non-trivial ones), in spite of the somewhat conservative nature of the stock examples of allowed variation. But if *that* is the case, then we can improve upon this strategy of conveying by example and in a piecemeal fashion what is meant by "categorical" by moving to a more general – and thus more definitive – characterization in the following way.

[8] Armstrong's example in fact concerns mass and Newton's law of gravity, but the claims are perfectly analogous.

Recall that the example just looked at was that of Coulomb's law. This law is a paradigmatic example of a *classical* law, and as such of a *functional* law. That is, Coulomb's law is a law of the form

$$a(x) = f(b(x), c(y), d(x, y)),$$

where $a(x)$, $b(x)$, $c(y)$ and $d(x, y)$ are real- (or real vector-) valued functions representing the determinable physical properties A, B, C and the relation D, and f is some *functional* (that is, a function of functions). Thus note that the template for laws that is assumed in the contemporary debate is *not* the old $\forall x(Fx \rightarrow Gx)$-type formulation that was so central to earlier discussions. The stated reason that Armstrong provides for this move away from the older representation is that

> The laws that have the best present claim to be fundamental are laws that link together certain classes of universals, in particular, certain determinate quantities falling under a common determinable, in some mathematical relation. They are functional laws. If we can give some plausible account of functional laws, then and only then do we have a theory of lawhood that can be taken really seriously (Armstrong 1993, p. 242).

Assuming such an account of fundamental laws, then, we can better formalize the canonical debate over their modal interpretation as follows. Suppose first of all that a fundamental law, say an actual fundamental law, is given by

$$a(x) = f(b(x), c(y), d(x, y))$$

for some specific properties and relations A to D. Anti-Humeans will then hold that since the fundamental properties are essentially dispositional, it follows that

$$\neg \diamond a(x) \neq f(b(x), c(y), d(x, y)),$$

and in particular that

$$\neg \diamond a(x) = f'(b(x), c(y), d(x, y))$$

for any $f' \neq f$. Thus in this context in which laws are conceived of in functional terms, it is not merely the *properties* to which a given property is related to that must be held fixed across possible worlds, but also the *way in which* it is so related, where that "way" is expressed in terms of a functional connection between them. By contrast, Humeans will hold that

$$\diamond a(x) \neq f(b(x), c(y), d(x, y)),$$

and in particular that

$$\diamond a(x) = f'(b(x), c(y), d(x, y)),$$

for at least some f'.

How widely should f' be allowed to vary? As noted above, if fundamental properties are categorical then it seems right to say that there should be *no* non-trivial constraints on the form of the laws that such properties feature in, and hence no non-trivial constraints on the choice of the functional f'.[9] But then another and more perspicuous way to characterize a categorical property is as one that is "independent of its nomic role" (Mumford 2004, p. 150), where – as we are now in a position to state – that role is *defined* by (i) the functional form of the law and (ii) the identities of the properties to which the property is functionally related. That, I take it, may be regarded as the sought-for precisification of what is meant by "categorical property", and as such of the Humean package as well.

That completes my outline of the canonical debate over the modal status of the laws of nature, as I understand it. What is assumed, first of all, is a fundamental modal distinction between properties which sorts them into "categorical" and "essentially dispositional", where I take the former sort of property to be most perspicuously defined as a moment ago. That modal distinction between properties is then used to ground a corresponding modal distinction between laws, defining the contingentists and necessitarians respectively. As is manifest from the quote from Armstrong above (and as would have been obvious anyway), it is explicit in this debate that the laws of nature of principal interest are the *fundamental* laws, and it is equally explicit that these laws are assumed to have a functional structure. But when the terms of the canonical debate are stated in that way, it becomes glaringly obvious that there is an urgent problem afoot in it. That problem, of course, is that this debate over laws played out in the metaphysics literature purports to describe fundamental laws and properties, and thus to capture the metaphysics of fundamental physics; *but fundamental physics properties do not obey functional laws*![10] The reason for this, of course, is that fundamental properties and the laws that relate them must be understood within the framework of quantum theory, and the laws of quantum systems simply cannot be shoe-horned into functional form.[11] But since categorical properties have been *defined* in terms of the relationship they bear to functional laws, what we need to consider now is the question of whether *any* fundamental

[9] By "trivial" constraints on the functional form of laws that a categorical property A can participate in, I have in mind general conditions such as (i) there is no A-dependence on the right-hand side that cancels the occurrence of A on the left (as in $a = f(b, c) + a$), or (ii) the form of the equation does not make it inapplicable to some of the determinates associated with the determinable (as in $a = 2$), etc.

[10] Or if we count charge as a fundamental property (which is controversial), at least not in its most "fundamental guise" – Coulomb's law is after all just an approximation to the laws of quantum electrodynamics.

[11] To cite just one reason: the fact that in quantum mechanics some properties are quantized and others continuous entails that properties in general can no longer be representable by continuous functions, but rather should represented by matrices (some of which have continuous spectra and others discrete).

property can be classified as categorical when the latter are out of the picture.[12] Let me therefore now consider whether the fundamental properties may be regarded as categorical in the context of contemporary physics, and thus whether categorical properties may still be appealed to in order to ground a contingentist interpretation of laws. As above, I will continue to focus on the fundamental kind properties.[13]

3 Laws and Properties in Modern Physics: Problems for Humeanism

Given that functional laws are now out of the picture, the first item on the agenda is to clarify how laws are in fact represented in modern fundamental physics contexts. Since quantum field theory (QFT) is the most fundamental physical framework we have developed to date – certainly if we restrict our attention to those we can subject to empirical test – it is the structure of laws in QFT that we would ideally attend to directly. However, most of the critical points in what follows can be stated in the (far simpler) context of quantum particle mechanics, and as such I will elect to do so whenever we can get away with it. (While quantum particle mechanics will serve us well in what follows, the final point I wish to make requires concepts in QFT specifically.)

To begin, then, let us focus on quantum particle mechanics. In this context, the nearest thing we have to a template for laws along the lines of the functional template of classical physics is the Schrödinger equation. Despite the presence of the definite article, "the" Schrödinger equation is not so much *an* equation as a structure into which the various laws of quantum particles must slot, and laws of Schrödinger form relate the evolution of such a system to the action of the relevant Hamiltonian on the system states. They are therefore statements of the form

$$i\hbar\frac{\partial|\psi(n_i)\rangle}{\partial t} = H_\alpha|\psi(n_i)\rangle,$$

where H_α denotes some specific Hamiltonian and the n_i denote the properties that identify the kind, or kinds, of particle involved.[14] These Hamiltonians describe both

[12] Since essentially dispositional properties are typically characterized in terms of their entailment of such laws, analogous problems will apply to them; but I forgo discussion of such properties here. The central point of this paper, after all, is to show that this whole debate needs to be rethought, not that some one side of it triumphs over the other.

[13] Since state-dependent properties in quantum mechanics are typically taken to be possessed only conditionally upon measurement, it is already clear that it will be difficult to maintain that *they* are categorical.

[14] Some state-dependent variables x_i should also be included in the characterization of the state, but as my focus is just on kind properties here I omit them in what follows.

how a single particle's states evolve in time through its Hilbert space, but also contain all the information about a particle's *interactions* with other systems. For example, the quantity

$$\langle (n, \pi^+) | H_S | (p, \pi^-) \rangle \tag{1}$$

yields the probability that two different particle kinds, here a negative pion and a proton, will interact through the strong interaction when smashed together in an accelerator to produce a positive pion and a neutron. These probabilities concerning which particles will be produced when others interact in this way essentially exhaust the empirical output of a theory of particle physics.

These facts about laws and probabilities in quantum mechanics are utterly elementary. But behind the elementary nature of these facts hide some important implications for modal metaphysics. The most expedient way to see this is to attend straight to the case in which there is some non-trivial *symmetry* at play in the dynamics expressed by H_α. While this is admittedly a special case in the space of all possible Hamiltonians, it is emphatically *not* a special case from the perspective of actual fundamental physics, since all the known fundamental interactions are associated with some significant symmetry. In the context of quantum particle mechanics, to say that a law exhibits a symmetry is to say that there exists a set of observables U_i such that (i) the U_i are the generators of a unitary Lie group, and as such define a Lie algebra, and (ii) for all U_i, $[H_\alpha, U_i] = 0$, where H_α is the Hamiltonian corresponding to that law and $[H_\alpha, U_i] = H_\alpha U_i - U_i H_\alpha$. The presence of such a symmetry has important consequences for the solutions of the Schrödinger equation (here presented in time-independent form) – namely, that

$$H_\alpha \psi(n_i) = E\psi(n_i) \Rightarrow H_\alpha(\psi(n_i')) = E\psi(n_i'),$$

where E represents energy, the n_i again represent a set of determinate properties defining some kind, and the n_j a different set of determinates *but of the same determinables* as those that define the first. Thus where there exist symmetries in the laws, there exist *families of particles* that obey those laws with the same energy (hence in relativistic contexts same mass), but that have different magnitudes of the same determinable properties. Such of families of particles are denoted as "multiplets". It is essentially because of this tight connection between symmetries and particle kinds that physicists are able to predict the existence of new particles long before they are seen in the lab (famous examples being the Z^0 and Ω^-).[15]

As already noted, the actual laws of physics themselves possess a great deal of symmetry: we have, for example, the SU(2)⊗U(1) symmetry of the electroweak

[15]It should be noted that in the (more complicated) context of QFT, it remains that both the fields and the field quanta (particles) that act in accordance with a law will form multiplets of whatever symmetry that law is taken to have. As such, this observation concerning how the distribution of particle property magnitudes relates to the presence of dynamical symmetries transfers directly to the (more fundamental) quantum field-theoretic context.

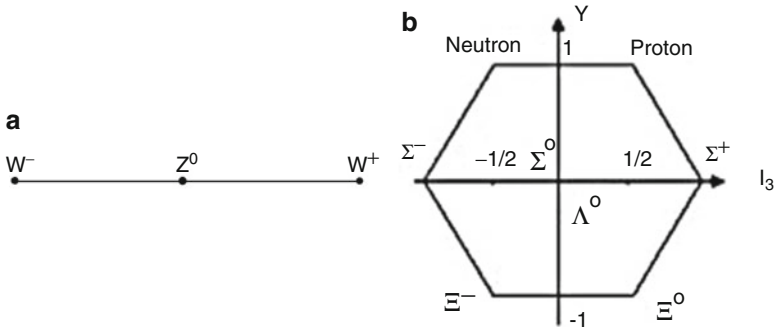

Fig. 1 Some actual particle multiplets. (**a**) SU(2) Triplet of weak bosons. (**b**) SU(3) octet of hadrons

interaction, and the SU(3) symmetry of the strong interaction. That means, of course, that the particles (and fields) that populate this world themselves fall into such multiplets. We have in Fig. 1a, for example, the triplet of the weak bosons defined by their differing values of weak isospin, corresponding to the 3-membered multiplet of the SU(2) symmetry in the electroweak interaction. In Fig. 1b, we find the neutron, the proton and other hadrons comprising and 8-membered multiplet of SU(3).[16]

3.1 Problems for Categoricalism

These diagrams represent elegant facts about the fundamental structure of the actual world. To see their principal relevance for topic at hand, however, it helps to recall that debates over the modal status of laws are often framed in terms of *duplicates*. We already know that Anti-Humeans hold that otherworldly duplicates of actual particles cannot act in accordance with different laws; as Bird puts it, for Anti-Humeans "if the particles and fields are the same in two worlds then they instantiate the same [essentially dispositional properties] and thus give rise to identical laws" (Bird op. cit., p. 84). Humeans of course deny this, holding that otherworldly duplicates of actual particles may accord with different laws (see e.g. Lewis 1986, p. 163). I have already pointed out that the extant literature is basically silent on the issue of just how much nomic variation is possible given categoricalism, but I have argued that the informal renderings of "categorical" point to the idea that it should be possible for such properties to feature in laws with an *arbitarily* different

[16]While hadrons are no longer regarded as fundamental particles, the gluons – which currently are so regarded – form a multiplet with an identical structure.

structure. What, then, is the situation here? Can the behaviour of otherworldly duplicates of the actual particles, which as we know occur in *multiplets*, be described by laws with a structure arbitrarily different from those of the actual laws?

But the answer to this question is a clear and resounding *no*. A little more technically, what the above diagrams represent are *weight diagrams* of the algebras corresponding to the relevant symmetry.[17] Each weight in a diagram corresponds to a member of a particle multiplet, and the operators of the algebra may be defined so as to map between the various weights in the diagram. Each diagram may therefore be thought of as a sort of solution space for the algebra. Furthermore, a little more attention to the mathematical theory describing these entities – namely, the theory of semi-simple Lie algebras – allows one to deduce that each such weight diagram corresponds to *one and only one algebra*.[18] What that informs us of in turn is that, wherever it is in possibility space that duplicates of these actual kinds are instantiated, then if we understand the laws operative there along quantum-mechanical lines it follows that those laws *must possess the symmetry of the laws of the actual world*. But that represents a *hugely* informative and non-trivial constraint on the laws that any such set of duplicates can accord with. It may be shown, for example, that the probabilities for kind production given above in (1) may essentially be computed through consideration of symmetry alone; since as already noted the probabilities associated with outcomes of interactions essentially exhaust empirical output of particle physics theories, information about symmetry is *highly* non-trivial information from an empirical point of view.[19] As a result,

[17] An algebra is one way to characterize a symmetry. It is not the most fine-grained way, granted, since the algebra is insensitive to the global properties the associated group describes. Nonetheless, it remains that to pin down a symmetry up to the level of the associated algebra is still to determine a highly non-trivial constraint, and here – just as in most contexts in particle physics – I will be content to describe symmetries in algebraic terms.

[18] This is at least the case for the compact semi-simple algebras, which are those typically employed in particle physics when dealing with internal symmetries. The basic idea behind this is as follows. The operators of any simple compact Lie algebra can be arranged into a maximal set of r commuting operators H_i, and a remaining set of ℓ operators E_α. Since the r commuting operators can be used to define those properties of the particles that can all be observed simultaneously with one another (and with the energy), it is the eigenvalues of these operators that are taken to define particle kinds. Each of these kinds is represented by a weight in the diagram. The remaining operators E_α are "step" operators that map from one weight (particle) to another weight displaced from the first by a vector α. These α are also the "roots" of the algebra, and the full set of these roots may be used to classify the algebra. It follows from the last point that two distinct algebras must differ in their roots, and thus differ in terms of how the weights in their associated weight diagrams are displaced from one another; it follows from that in turn that each weight diagram corresponds to one algebra only. And since semi-simple Lie algebras are just the sums of simple Lie algebras, each set of particles associated with *these* algebras will likewise correspond to a unique such algebra.

[19] It should be noted that we must in general know more than the relevant internal symmetry group to predict the outcome of scattering experiments, as these outcomes are distributions in space (so that information regarding the external symmetries, the structure functions describing the material constitution of the colliding particles, etc., has to be invoked in addition).

knowledge of the symmetry associated with an interaction is in most cases the single most significant piece of information from a modern physicist's point of view, and indeed entire research programmes have flourished in the absence of any knowledge concerning the relevant laws that transcended their symmetry properties alone.[20]

But now we are in a position to see why the appeal to categorical properties is so very problematic in the fundamental physics context. We saw that in the canonical account – at least as I reconstructed it – the definitive feature of categorical properties was their *failure to place any non-trivial constraints* on the laws they feature in, including the structural *form* of those laws. What we see now, however, is that if we conceptualize fundamental laws in a way that approximates how physics in fact understands them, then in any possible world in which the actual kinds are reduplicated the symmetries of the actual laws must be reduplicated as well. But we know that this represents a *highly* non-trivial structural constraint on the laws that particle kinds can satisfy. How, then, can we say that the associated kind properties are categorical, if *part of what it is* to be categorical is to be free of such constraints?

The appropriate conclusion to draw at this point therefore seems to be this: we simply cannot understand the fundamental properties as categorical in character, at least not in anything like the sense in which such properties are canonically understood. And since it is primarily in the works of the canonical debate that these properties are articulated, it therefore isn't clear that there is any hope of retaining categoricalism at all.

3.2 Problems for Contingentism

The above considerations strongly suggest that the Humean must relinquish their commitment to categoricalism regarding the fundamental kind properties. But given that – in the canonical account at least – their modal interpretation of laws is *grounded* in the prior commitment to these properties, the same considerations surely threaten the contingentist approach to laws that defines the second aspect of their view. What I want to argue now is that these considerations do indeed put pressure not just upon the Humean interpretation of properties, but on their interpretation of laws as well.

The reason why the above considerations pose a threat to nomic contingentism should be immediately clear. The problems for categoricalism outlined above derive from the fact that the fundamental kind properties of this world, when conceived of post-classically, impose significant structural constraints on the laws that those kinds can satisfy. As such, the tighter these constraints get, the more diminished the scope for contingency. We saw above that these constraints amount to the fact that reduplicating the actual particles and fields reduplicates the symmetry associated

[20]The Eightfold Way programme of Gell-Mann and Ne'eman is the classic example here (see Gell-Mann and Ne'emann 1964).

with the actual laws, so what we must contemplate in order to go on is the extent to which pinning down a law's symmetry sufficies to pin down the law itself. Let us therefore attend to that now.

It was registered above that identifying the symmetry associated with a law furnishes us with highly non-trivial knowledge, from both a theoretical and an empirical point of view. However, it is not the case in general – either in classical, quantum particle, or quantum field theory – that a law is uniquely determined by the symmetry associated with it; knowing that a classical electromagnetic potential is spherically symmetric, for example, will not be enough to pin down the radial dependence of the potential uniquely (see e.g. Martin 2003 for discussion). However, as was mentioned above, the protagonists in this debate are not primarily concerned with natural laws *in general*: rather, they are primarily interested in giving a metaphysics of the truly *fundamental* laws and properties. The question we ought to focus on, then, is that of whether the *those* laws may be determined by symmetries alone.

As things turn out, and very interestingly, there is at least a case to be made that we should answer that question in the affirmative – for one can argue that the symmetry associated with a fundamental law *does* in fact suffice to determine it uniquely.[21] This is due to the fact that fundamental laws are required to have properties that less fundamental laws arguably need not, and the reason for this is roughly as follows. Again, the fundamental laws, if such there be, must be described in the most fundamental framework we have, and at the moment that is QFT. It has been understood since its earliest days that quantum field theories are plagued with divergences that must be removed through the complicated and arduous process of renormalization.[22] It came to light only later, however, that we can expect even properly renormalized theories to diverge at sufficiently high energy. Since relative to the background of QFT's assumptions a theory must be valid to arbitrarily high energy if it is to give a fundamental description of nature, of considerable interest to physicists is what properties a QFT must have in order to be valid in this limit.[23] In this connection, the physicist Frank Wilczek has argued that the only quantum field-theoretic laws that can be shown to exist in a computationally tractable way in this infinite-energy limit are those that are "asymptotically free" – that is, those whose couplings disappear at infinity (Wilczek 1999). Furthermore, Wilczek and others have shown that the only asymptotically free quantum field theories that can exist in four dimensions are the so-called *renormalizable local gauge theories* (see, e.g., Gross and Wilczek 1973; Politzer 1973). But the pertinent point about such theories is that, for a given matter content, these are essentially *uniquely specified* once we

[21]Modulo the assumption that we already have the kind ontology that the laws are supposed to describe in place, which in the context of particle reduplication scenarios will obviously be the case.

[22]Any introductory QFT textbook will outline this process for the uninitiated.

[23]The requirement that QFTs must be regarded as valid in the infinite energy limit follows from the spacetime continuity required by Lorentz invariance in tandem with the complementarity of 4-momentum and spacetime in quantum mechanics.

have specified the relevant symmetry. Putting things a little more precisely, on the assumption that the fields concerned are specified, the laws are thereby also uniquely specified *but for the values of the constants* appearing in them. Determination of these constants is therefore a matter of matching them to experiment.

What we see, then, is that if we want to give a fundamental modal metaphysics that reflects current fundamental physics, there are compelling reasons to reject the claim that the actual particles and fields could behave very differently from the way that they actually do. This is because there are reasons to hold that the only variable factor in the laws describing their behaviour is the values of the numerical constants featuring in them. Now, it should be underlined that such differences in the values of constants can make for profound physical differences between *worlds*: think, for example, of how different this world would be if the electromagnetic interaction were a 100 times stronger at the distance of a femtometer, so that the strong force keeping the protons in atomic nuclei together was overtaken by the electromagnetic repulsion pushing them apart (see Quigg 2007). Nevertheless, it remains that the degree of *nomic* variation permitted by the above-cited results is radically diminished in comparison to that countenanced in the canonical account. Indeed, if change in the values of constants is the extent of variation that the nomic contingentist can lay claim to, I submit that they are barely entitled to call themselves nomic contingentists at all.

4 Coda on Humeanism

The considerations outlined above regarding particles, symmetries and laws in modern physics demonstrate that each component of the Humean package has been either ruled out or, at the very least, backed into a tiny corner. It may therefore seem that we can only conclude that the Humean stance towards laws and properties is simply dead in the water. In particular, the fact that we are now encroaching upon necessitarianism about fundamental laws seems to be in flat contradiction with the Humean dictum that there are no "necessities in nature", or at least no necessities that cannot be reduced to those of mathematics or logic. However, by way of rounding off this discussion I beg that we reconsider whether the necessities gestured at above are as unpalatable to Humeans as they may at first appear. I will argue that they are *not*, and as such that a Humean approach to laws is very much still in the offing.

At the root of the problems outlined above is the fact that, in quantum frameworks, the kind content of a given world goes a long way to determining the symmetry structure of the laws in that world.[24] The reason for that in turn is that, assuming that the laws concerned exhibit symmetries, the particles acting in

[24]I was explicit above that I am making the assumption above that the worlds and laws in question exhibit some non-trivial symmetry. If this assumption is dropped my argument does not get off

accordance with those laws will fall automatically into *multiplets*, and the governing theory of Lie algebras dictates that each such multiplet corresponds to one and only one symmetry. But what I then ask that we consider is this: if this fact about the relation between Lie algebras and the associated multiplets is what lies at the root of the necessitation problem, then *what exactly is it* about this necessitation that is unacceptable to Humeans? Hume himself, after all, was perfectly happy to countenance the existence of necessities in the realm of "relations of ideas", and so in algebra, arithmetic and geometry; but is it not at bottom a *mathematical* fact that a set of particles, defined by a given set of determinate values, cannot participate in laws of quantum-theoretic form with arbitrary symmetry structure? It seems to me that it is; and as such, it seems to me that although the laws describing a given set of particles may be unique and in that sense necessary, it is not a necessity that Hume himself need have felt particularly troubled by. And if the protagonists in the canonical debate had only realized that they were working with an unacceptably classical account of laws of nature for which the above considerations regarding kinds cannot even get off the ground, they too may have realized that Humeans can in fact countenance nomic necessity of this sort.

At this point I envisage an immediate objection to the above remarks concerning the necessity of the laws of nature and their compatability with Humeanism. This objection is that my argument suggesting that the laws that, for example, the hadrons can accord with are unique and in that sense necessary was made *relative to the assumption* that those laws are conceived of quantum-field theoretically. But since, one might hazard, there could be a world in which there are hadrons but in which *classical* physics holds sway, the necessitation arrived at is relative to that (substantive) assumption and is in that sense at best a "contingent" necessity.[25] There are three things I would like to say by way of a response to this. Firstly, it is in fact entirely unclear that there could be a world that is both fundamentally classical and that could serve as an object of study by physics (see Mirman 1995, Chap. 1). Secondly, the canonical debate likewise makes substantive and contingent assumptions when it assumes that laws can be fit into the functional template. Therefore if the objection to my conclusion that the fundamental laws are necessary is that there could exist worlds in which the laws cannot even be expressed in terms of the relevant quantum-theoretic template, then one could make an exactly parallel objection to the Anti-Humeans in the canonical account – and indeed against anyone who would claim that any law at all is necessary. For how do *they* know what possible templates for laws are out there in possibility space? However, it seems to me that if we are to argue about whether the laws are necessary or not, *we need to agree at the outset* as to what sort of structural templates are going to count as laws *and then argue within that assumption*, by arguing about what variations within

---

the ground, though as pointed out above making the assumption does not entail that the cases concerned are particularly "special" from a physics point of view.

[25]Here I do not wish to connote the "contingent necessity" associated with Armstrong's analysis of laws.

that template are possible. Finally, however, insofar as the debate concerns what duplicates of a given set of entities are capable of in other possible worlds, it seems clear to me that we need to settle at the outset *what those entities are*. If we agree that those entities are quantum entities of some sort (whether particles or fields), then we *have no choice* but to use laws of the appropriate quantum form. As such, there really *is* no contingency in the framework we adopt to describe the behaviour of duplicates of the posited entities, and the above objection is simply moot.

5 Conclusion

I will be the first to admit that the argument I have laid out above is very sketchy in many places – far sketchier than it would have to be to establish its somewhat substantive conclusion. Nevertheless, I hope that I have, managed to convey that the modal disputations that we routinely engage in scientific metaphysics need to change, and change dramatically. I hope to have shown, first of all, that the "canonical" debate played out in the metaphysics literature is explicitly wedded to an unacceptably classical view of the world. I hope furthermore to have shown that when we try to reproduce the debate in the context of laws that bear more similarity to the fundamental laws of physics, we can no longer claim that the fundamental kind properties are "free of nomic implications", and as such categorical in character; nor is there much scope to argue that the fundamental laws are contingent. However, I have argued that this necessaritarianism might, contrary to appearances, be entirely palatable to Humeans: it will be a Humeanism that is not built on the ediface of categorical properties, granted, but I think we can call it a Humeanism about laws nonetheless. Given that necessitarianism about laws was partly definitive of the *Anti*-Humean position in the canonical debate, this circumstance vividly suggests to me that paying attention to the mathematics of real physics can change the landscape of modal metaphysics quite dramatically. What is crystal clear to me, in any case, is that we cannot continue to kid ourselves into thinking that by contemplating only the possibilities that might be sanctioned classically we can thereby arrive at a modal metaphysics that has any right to be called "fundamental".

References

Armstrong, D.M. 1983. *What is a law of nature?* Cambridge: Cambridge University Press.
Armstrong, D.M. 1993. *A world of states of affairs*. Cambridge: Cambridge University Press.
Armstrong, D.M. 2005. Four disputes about properties. *Synthese* 144(3): 309–320.
Bird, A. 2007. *Nature's metaphysics*. Oxford: Oxford University Press.
Cohen, J., and C. Callender. 2009. A better best system account of lawhood. *Philosophical Studies* 145: 1–43.
Carroll, J.W. 1994. *Laws of nature*. Cambridge: Cambridge University Press.

Earman, J., and J. Roberts. 2005. Contact with the nomic: A challenge for deniers of humean supervenience about laws of nature (Part I). *Philosophy and Phenomenological Research* 71: 1–22.

Ellis, B. 2001. *Scientific essentialism*. Cambridge: Cambridge University Press.

Gell-Mann, M., and Y. Ne'emann. 1964. *The eightfold way*. New York: Westview Press.

Gross, D., and F. Wilczek. 1973. Asympotically free Gauge theories. *Physics Review Letters* 30: 1343.

Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy* 61(4): 343–377.

Lewis, D. 1986. *On the plurality of worlds*. Oxford: Blackwell.

Loewer, B. 1996. Humean supervenience. *Philosophical Topics* 24: 101–127.

Martin, C. 2003. On the continuous symmetries and the foundations of modern physics. In *Symmetries in physics*, ed. K. Brading and E. Castellani, 29–60. Oxford: Oxford University Press.

Mellor, H. 2000. The semantics and ontology of dispositions. *Mind* 109(436): 757–780.

Mirman, R. 1995. *Group theoretical foundations of quantum mechanics*. Commack: Nova Science Publishers.

Mellor. 2003. Real metaphysics: Replies. In *Real metaphysics: Essays in honour of D.H. Mellor*, ed. H. Lillehammer and G. Rodriguez-Pereyra, 212–238. London: Routledge.

Mumford, S. 1998. *Dispositions*. Oxford: Oxford University Press.

Mumford, S. 2004. *Laws in nature*. London: Routledge.

Politzer, H. 1973. Reliable perturbative results for strong interactions? *Physical Review Letters* 30: 1346–1349.

Quigg, C. 2007. Spontaneous symmetry breaking as a basis of particle mass. *Reports on Progress in Physics* 70: 1019–1054.

Wilczek, F. 1999. What QCD tells us about nature – and why we should listen. arXiv:hep-ph/9907340.

The Undeniable Effectiveness of Mathematics in the Special Sciences

Mark Colyvan

1 The Philosophical Problems of Applied Mathematics

The applications of mathematics to empirical science raise a number of interesting philosophical issues. Perhaps the most well known of these issues is the so-called unreasonable effectiveness of mathematics. The issue here is to account for the success of mathematics in helping empirical science achieve its goals. It is hard to say precisely what the crux of the issue is supposed to be, let alone what an adequate explanation would look like. The problem is usually attributed to Eugene Wigner (1960) in his well known essay on the topic,[1] where he suggests that

> [t]he miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. (Wigner 1960, p. 14)

I take it that the problem, in its most general form, is to account for the applicability of mathematics in empirical science. Put this way, though, there are a number, of interrelated problems. There's the unreasonable effectiveness of arithmetic, of calculus, of differential geometry, of algebraic topology, and so on.[2]

[1] There has been a great deal of subsequent discussion on the issue, for example Azzouni (2000), Colyvan (2001b), Grattan-Guinness (2008), Hamming (1980), Steiner (1998) and Wilson (2000) and this discussion has helped clarify the problem and its solution.

[2] Not to mention the much less appreciated problem of the unreasonable effectiveness of inconsistent mathematics (Colyvan 2009).

M. Colyvan (✉)
Department of Philosophy, University of Sydney, Sydney, NSW 2006, Australia

Stellenbosch Institute for Advanced Study (STIAS), Wallenberg Research Centre at Stellenbosch University, Stellenbosch, South Africa
e-mail: mark.colyvan@sydney.edu.au

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_5,

There's the way different philosophies of mathematics draw different conclusions to help explain the applications of mathematics.[3] There's the issue of the different roles mathematics can play in science – the different ways mathematics might be thought to be unreasonably effective. And, of course, physics is not the only scientific consumer of mathematics. Mathematics might also be thought to be unreasonably effectiveness in economics, in biology, in chemistry, in psychology, and elsewhere. Finally, there's the problem of understanding the nature of the modelling process itself and why mathematical modelling is so often an effective way of advancing our knowledge.[4]

Many of these issues are interrelated but, still, a great deal of confusion has resulted from running some of the issues together and failing to state exactly what is supposed to be unreasonable about the effectiveness in question. Having been guilty myself of such carelessness in the past (Colyvan 1999, p. 15), my aim here is a modest one. I intend to look at the use of mathematical models in the special sciences. As my primary example I'll consider the use of mathematics in population ecology. The issue here is that the mathematical models in question seem to leave out the relevant causal detail, yet still manage to both predict and (arguably) explain population-level phenomena. The task, then, is to give an account of how mathematical models can succeed in such tasks.

2 Case Study: Population Ecology

Population ecology is the study of population abundance and how this changes over time. For present purposes, a population can be thought of as a collection of individuals of the same species, inhabiting the same region. Population ecology is a high-level special science, but relies heavily on mathematical models. (It is thus a soft science in one sense – in the sense of being high level and quite removed from physics – but in another sense it is a hard science – in the sense that it is mathematically sophisticated.) There are a number of issues associated with applying mathematics to population ecology, but my focus here will be on an issue that is of significance for working ecologists and has a direct bearing on the way they go about their business.[5] The issue I will address arises from the fact that mathematical models apparently ignore the relevant biology and would thus seem ill-equipped to offer explanations of ecological phenomena. Mathematical models in population ecology would thus seem to be (at best) predictive models. In what follows I will argue that this is not right. I will argue that while, in a sense, mathematical models ignore the relevant biology, this does not mean that these

[3] See, for example, Colyvan (2001a), Frege (1970) and Steiner (1998).

[4] See, for example, Batterman (2010) and Bueno and Colyvan (2011).

[5] See for example Levins (1966) and May (2004) for some discussion on this and related issues by prominent ecologists.

models cannot be explanatory. I will also provide a sketch of how mathematics can succeed in delivering explanations, despite turning away from much of the biological causal detail.

Before I begin the main task, however, it will be useful to present a couple of typical mathematical models, of the kind we are interested in here. First consider *the logistic equation* (Gotelli 2001, Chap. 2). This is a model of a single population's abundance, N – exponential at first and then flattening out as it approaches carrying capacity, K:

$$\frac{dN}{dt} = rN(1 - \frac{N}{K})$$

where r, is the population growth rate and t is time.

Another key example is *the Lotka-Volterra equations* (Boyce and DiPrima 1986, Chap. 9) These equations model the population of the predator and the prey via two coupled first-order differential equations:

$$\frac{dV}{dt} = rV - \alpha VP$$
$$\frac{dP}{dt} = \beta VP - qP$$

Here V is the population of the prey, P is the population of the predator, r is the intrinsic rate of increase in prey population, q is the per capita death rate of the predator population, and α and β are parameters: the capture efficiency and the conversion efficiency, respectively. These equations can give rise to complex dynamics, but the dual out-of-phases, population oscillations of predator and prey are the best known.

Of course both these mathematical models are overly simple and are rarely used beyond introductory texts in population ecology. For example, the logistic equation treats the carrying capacity as constant, and the Lotka-Volterra equations treats the predators as specialists, incapable of eating anything other than the prey in question. Both these assumptions are typically false. These models do, however, serve as the basis for many of the more realistic models used in population ecology. The more serious models add complications such as age structure, variable growth rates and the like. These complications do not matter for my purposes in this paper, though. Even in these more complicated models, biological detail is deliberately omitted and yet the models are adequate for the purposes at hand. The issues I am interested to explore can be raised with the more complicated models, but it's easier to see the issues in the simpler models. We will not be losing any generality by focussing our attention on the simpler text-book population models.

We are now in a position to state the philosophical problem posed by mathematics in population ecology. Population abundance is completely determined by biological facts at the organism level – births, deaths, immigration and emigration – but the (standard) mathematical models leave out all the biological detail of which

individuals are dying (and why), which are immigrating (and why), and so on. That is, the mathematical models ignore the only things that matter, namely, the biological facts. The mathematical models here – the relevant differential equations – seem to ignore the biology, and yet it is the biology that fully determines population abundances. How can ignoring that which is most important ever be a good strategy?

We might put the point in terms of explanation: the mathematical models are not explanatory because they ignore the causal detail. The model may tell us that the abundance of some population at time t is N, but without knowing anything about the organism-level biology, we will not know *why* the population at time t is N and will have little confidence in such predictions. A full account of the relevant biology, on the other hand, would include all the causal detail and *would* provide the required explanations. Let's focus on this explanatory version of the puzzle because I think it is what underwrites the less-specific worries expressed in the previous paragraph.

Before I go any further, it will be useful to say a few words about explanation and philosophical theories of explanation. First, I take it that we simply cannot deny that there are population-level explanations in ecology. To deny this would, in effect, amount to giving up on explanation in the special sciences. Unlike physics, in the special sciences we do not have the option of reserving all genuine explanation for the fundamental level (or the fundamental laws). So the issue we are meant to be addressing is not that there can be no explanation in the special sciences. Rather, we take it for granted that there are explanations in the special sciences but that the mathematical models used in special sciences such as population ecology can not deliver explanations.

Next we might reasonably ask for a philosophical account of explanation, so that we are all on the same page. But that turns out to be difficult for a number of reasons, not least of which is that there is no generally-accepted philosophical account of scientific explanation. So, for present purposes, I shall be rather liberal about what counts as an explanation. I suggest that an intuitive understanding of an explanation as an answer to a "why question" will do.[6] It is important to keep in mind that explanation should not be confused with a more limited class of explanation known as *causal explanation*. There is no denying that causal explanation – tracing the relevant causal history of an event of interest – is one kind of explanation. I deny, however, that this is the only kind of explanation.[7] Explanations must be enlightening, and that's about all we really need to assume here.

[6]I also take an explanation to be that which is accepted as such in the relevant scientific community. This, of course, is not a philosophical account of explanation; it's just a constraint that I take very seriously. I think philosophical accounts of explanation need to (largely) agree with scientific uses of the notion of explanation. A philosophical account of explanation that does violence to scientific practice is of little interest to we naturalistic philosophers.

[7]It would take us too far afield to argue for this here, but see, Colyvan (2001a, Chap. 3), Ruben (1990), Smart (1990), and Sober (1983).

3 The Role of Mathematics

Now I turn to the task of investigating what makes these mathematical models in ecology tick. I will argue that there is no reason to suggest that mathematical models in ecology are not explanatory. I will suggest three different ways in which the models in question can explain. First, the mathematical models do not ignore the biological detail – at least sometimes the models in question are offering biological explanations, albeit explanations couched in mathematical terms. Second, understanding a system often does involve ignoring, or rather, abstracting away from, causal detail in order to get the right perspective on it. Finally, I'll suggest that mathematics can offer explanation for empirical phenomena.

Recall that we started out with the charge that mathematical models leave out all the relevant biological detail. But this is not quite right. Often the mathematical model is just representing the biology in a mathematical form. For example, in the logistic equation, all the information about births, deaths, immigration and emigration is packed into r and all the information about the resources is packed into the constant K. The information about the predators' impact on the per capita growth rate of the prey is summarised in the Lotka-Volterra equation in α – the capture efficiency parameter – and the information about the predators' ability to turn prey into per capita growth of the predator population is summarised by β – the conversion efficiency parameter. You might have misgivings about the representation of this information,[8] but this is a different objection. It's now a concern about the simplicity of the model. As I mentioned before, we can provide more complex models that relinquish some of the more unrealistic idealisations. These more complex models also have their idealisations, though. Indeed, it is part of the very enterprise of modelling that some details are ignored. So the basic concern about biological detail not being represented in the mathematical models under consideration is misplaced. Of course not all the biological detail is present in the model, but the fact remains that many of the key terms of the mathematical models have natural biological interpretations, or at least are representing or summarising the biological information in mathematical form. The mathematical models have a lot more biology represented in them than is typically appreciated.

In cases where the biology is represented in mathematical form, the model is indeed capable of offering perfectly legitimate biological explanations. For instance, think of the standard story of how population cycles arise as a result of predator–prey interactions. The cycles in question are solutions to the coupled differential equations in question (Boyce and DiPrima 1986, Chap. 9) but there is also a very natural biological explanation that can be extracted from the mathematical model: when the predator population is high the predators catch many of the prey so that the latter's population falls, but then there is less food for the predators, so

[8] You might, for example, object that r and K are represented in the logistic model as constants.

after a time the predator population also falls; but now there is less pressure on the prey population, so it recovers and this, in turn, supports an increase in the predator population (after a similar time lag). This cyclic behaviour falls out of the mathematics, but the explanation, once suitably interpreted, is in fact a perfectly respectable ecological explanation.

Next, notice that ignoring some detail can lead to insights via analogy.[9] Sometimes similarities between systems will not be apparent until certain details are ignored. Mathematics is a particularly useful tool for drawing out such similarities because mathematics allows one – indeed forces one – to abstract away from the causal detail and notice abstract similarities. For example, Newton's law of cooling/heating is just the logistic equation with abundance replaced with temperature of the body in question, and carrying capacity replaced with ambient room temperature.[10] Why are such connections between systems important? One reason is that it saves work: one can import results already at hand from work done elsewhere. Once the connection between the logistic equation and the cooling/heating equation are recognised, results from either area can be used by the other area (suitably interpreted, of course). Moreover, these rather abstract connections – often only apparent via the mathematics – can lead to new developments and, as we'll see shortly, even help with explanations.

We have already seen that mathematics can be the vehicle for delivering biological explanations, but often the mathematics can facilitate more transparent explanations. Mathematical models can sometimes do more than just represent the biology in mathematical form and then deliver essentially biological explanations of biological facts (albeit in mathematical guise). Sometimes the mathematics delivers explanations that would not be apparent otherwise. For example, the explanation of the different kinds of complex behaviour a population can exhibit as it approaches its carrying capacity – damped oscillations, asymptotic approach, overshooting and crashes – may be best seen via the mathematics of the logistic equation.

Finally, and most controversially, I'll argue that there can be genuinely mathematical explanations of empirical facts. Alan Baker (2005, 2009), Aidan Lyon and Colyvan (2008) and I (Colyvan 2001a, 2002, 2010) have argued that mathematical models can provide genuinely mathematical explanations of biological facts. A couple of much-discussed examples from the literature on this topic will help. Consider the question of why hive-bee honeycomb has a hexagonal structure. The answer, it turns out, is because of the honeycomb theorem (Hales 2001): a hexagonal grid represents the most efficient way to divide a surface into regions of equal area with the least total perimeter of cells (Lyon and Colyvan 2008). There are some biological and pragmatic assumptions required for this explanation to succeed. These include the assumption that bees have a limited supply of wax and need to

[9] See Colyvan and Ginzburg (2010) for more on analogical reasoning in ecology.

[10] And, as Ginzburg and I have argued elsewhere, the inertial view of population growth is mathematical similar to celestial mechanics (they both employ the same second-order differential equations) (Colyvan and Ginzburg 2003; Ginzburg and Colyvan 2004).

conserve it while maximising honey storage space. They also need to do this while still being able gain access to the hive from the outside.[11] But with these assumptions in place, the important part of the explanation seems to be purely mathematical and is provided by the honeycomb theorem. Any purely biological explanation will be too specific – it will tell the story of how one particular group of bees built one particular hive with a hexagonal structure – and will miss the general point that all hives built under such constraints *must* have a hexagonal structure. The hexagonal structure is a solution to an evolutionary optimisation problem and as such is not a mere accident of any particular hive construction.

Alan Baker (2005) offers an ecological example of a mathematical explanation. Baker considers why a particular species of North American cicadas have life cycles which are prime numbers: 13 and 17 years. The explanation of this surprising ecological fact is provided by number theory: having a prime number life cycle is a good strategy for avoiding predators. With a sufficiently large prime cycle any predators with similar life cycles will very rarely coincide with the most vulnerable stage of the cicada life cycle. It is also interesting to note that the two known cases of this phenomenon yield consecutive prime numbers – 13 and 17 – as the life cycles in question. This suggests that larger primes such as 19, 23, and so on, are impractical for biological reasons. And the smaller primes of 5, 7, and 11 leave the cicadas open to predators with life cycles of 10 years (as well as to predators with life cycles of 15 and 20 years), 14 years, and 22 years respectively. Again it looks like the mathematics – in this case elementary number theory – is carrying the bulk of the explanatory load here.

One final example of a mathematical explanation in ecology. Here I will also illustrate how analogical reasoning can play an important role in delivering the mathematical explanation. As I noted earlier, populations cycles are one of the more well-known solutions of the Lotka-Volterra equations, but there are other, more general models of population cycles. The more general models invoke a second-order differential equation (instead of the coupled first-order equations in the Lotka-Volterra model) and allow for single-species population cycles (Ginzburg and Colyvan 2004). This more general approach to population cycles is mathematically very similar to two-body problems in celestial mechanics, with its periodic solutions to two-body problems.[12] This interdisciplinary connection is interesting in its own right but it is much more than a mere curiosity. This analogy has the potential to drive a number of developments in population ecology. First, the similar mathematical treatment suggests that there ought to be an ecological counterpart of inertia in physics, and this has lead to investigations into "ecological inertia" (essentially cross-generational time lags in population responses to changes in environment) (Inchausti and Ginzburg 2009).

[11] Hence the problem is a tiling problem and not a sphere-packing problem.

[12] Hence the phrase "ecological orbits", and the analogy of population cycles in ecology with planetary orbits (Ginzburg and Colyvan 2004).

A second development arising from the analogy in question is that there should be stable and unstable orbits, as is the case with satellite orbits. In the rings of Saturn, for instance, there are well-defined gaps marking out the unstable orbits of this system. Similarly, in the asteroid belt between Mars and Jupiter there are gaps – the Kirkwood gaps – and these represent unstable orbits as a result of resonance effects with other massive bodies (most notably Jupiter). One might well expect to see similar gaps in population cycles (Ginzburg and Colyvan 2004, pp. 52–57) and these gaps, if they exist, would be explained mathematically, by appeal to very general structural features of the systems in question (essentially by an eigen-analysis). Not only would such explanations be mathematical, they would have been discovered by way of an analogy, facilitated by the mathematics in question.

If Baker, Lyon and I are right about such cases being cases of mathematics carrying the bulk of the explanatory load, there is still an interesting question concerning how mathematics can do this. There are several possibilities here:

(i) Mathematics can demonstrate how something surprising is possible (e.g. stable two-species population cycles).
(ii) Mathematics can show that under a broad range of conditions, something initially surprising must occur (e.g. hexagonal structure in honeycomb).
(iii) Mathematics can demonstrate structural constraints on the system, thus delivering impossibility results (e.g. certain population abundance cycles are impossible).
(iv) Mathematics can demonstrate structural similarities between systems (e.g. missing population periods and the gaps in the rings of Saturn).

If all this is right, it is simply a mistake to assume that because mathematical models ignore some of the biological detail they are not capable of delivering explanations. Indeed, to deliver the explanation in at least some of these cases might *require* that some biological detail be ignored.[13] Given the modal character of the three kinds of explanation just mentioned (involving possibility, necessity, and impossibility), it is hard to see how any causal explanation can deliver such explanations.

4 A Cure for Physics Envy

Let me finish with a word of caution, lest I be accused of "physics envy". Physics envy is the intellectual crime of being over impressed with the technical, theoretical accomplishments of physics and trying to shoehorn ecology into more sophisticated mathematical treatments than are warranted by ecological data and theory (Cohen 1971; Egler 1986). The mistake in question is not a mistake of using mathematics at all in ecology; it's the mistake of using inappropriate – and, in

[13] See Batterman (2010) for more on the role of abstraction in such explanations.

particular, overly-complicated, inappropriate, and physics-inspired mathematics – in ecology.[14] Rather than being guilty of physics envy, I have been attempting to offer a cure for it – or at least offer something to ease some of the associated discomfort it brings on. I have argued that in at least some cases importing mathematical models from physics and liberating mathematical models in ecology of some of the biological detail can genuinely advance ecology. I am not advocating mathematics for mathematics sake (at least not here). Great skill is required to use mathematics in ecology in such a way to enlighten and not obscure.[15] All I have argued is that when used effectively, mathematics can play a number of important roles in ecological theory. Moreover, the full range of these roles has not been fully appreciated in at least the philosophical literature on the applications of mathematics. Once the roles of mathematics in the special sciences are better appreciated and understood – especially the explanatory roles – the effectiveness of mathematics seems less unreasonable.

I have argued that mathematics can play a number of useful roles in ecological theory. Mathematics can represent biological facts and it is often able to do this is such a way as to make certain biological explanations more accessible. Mathematics is well suited to drawing attention to similarities between apparently different systems (and often provide the appropriate level of abstract representation for investigating the similarities). This, allows each of these areas to learn from one another, and reduces duplication of research. Finally, I argued that there are explanations in ecology where the mathematics carries the bulk of the explanatory burden, and these explanations are appropriately seen as mathematical explanations of biological phenomena.

Although I have focussed on ecology as my primary case study, I suspect that much of what I have said will carry over fairly straightforwardly to at least some other special sciences. In particular, similar debates about the role of mathematical models, physics envy, and the like can be found in economics (Mirowski 1989). Not surprisingly, very similar models are employed in both economics (especially macroeconomics) and ecology (since both have exponential growth and decline as a fundamental assumption in their respective dynamics) so the generalisation to economics is not much of a stretch at all. Casting my net wider to other special sciences is not so straightforward, although I do expect similar stories, albeit with

[14]It is interesting to note that the pioneering work on population ecology, conducted independently by Lotka and Volterra, demonstrated quite different attitudes towards mathematization. Lotka was more inclined to import mathematics from physics and to invoke analogies to motivate such importation (Kingsland 1985; Israel 1988). While it would be unfair to charge Volterra with physics envy, still he seemed to have had no hesitation in adopting the mathematical methods of physics when developing ecological theories. In this sense, the debate over physics envy might well be traced back to differences in the methods of the two founding fathers of population ecology.

[15]See May (2004) for more on this.

quite different mathematical models in the spotlight. For now, however, I am content if I have illuminated the applications of mathematics in one special science, namely population ecology.[16]

References

Azzouni, J. 2000. Applying mathematics: An attempt to design a philosophical problem. *Monist* 83(2): 209–227.

Baker, A. 2005. Are there genuine mathematical explanations of physical phenomena? *Mind* 114(454): 223–238.

Baker, A. 2009. Mathematical explanation in science. *The British Journal for the Philosophy of Science* 60: 611–633.

Batterman, R.W. 2010. On the explanatory role of mathematics in empirical science. *The British Journal of Philosophy of Science* 61(1): 1–25.

Boyce, W.E., and R.C. DiPrima. 1986. *Elementary differential equations and boundary value problems*, 4th ed. New York: Wiley.

Bueno, O., and M. Colyvan. 2011. An inferential conception of the application of mathematics. *Noûs* 45(2): 345–374.

Cohen, J.E. 1971. Mathematics as metaphor. *Science* 172(14 May): 674–675.

Colyvan, M. 1999. Confirmation theory and indispensability. *Philosophical Studies* 19(1): 1–19.

Colyvan, M. 2001a. *The indispensability of mathematics*. New York: Oxford University Press.

Colyvan, M. 2001b. The miracle of applied mathematics. *Synthese* 127(3): 265–278.

Colyvan, M. 2002. Mathematics and aesthetic considerations in science. *Mind* 111(441): 69–74.

Colyvan, M. 2009. Applying inconsistent mathematics. In *New waves in philosophy of mathematics*, ed. O. Bueno and Ø. Linnebo, 160–172. Basingstoke: Palgrave MacMillan. Repr., In *The best writing on mathematics 2010*, ed. M. Pitici (2011), 346–357. Princeton: Princeton University Press.

Colyvan, M. 2010. There is no easy road to nominalism. *Mind* 119(474): 285–306.

Colyvan, M., and L.R. Ginzburg. 2003. The galilean turn in ecology. *Biology and Philosophy* 18(3): 401–414.

Colyvan, M., and L.R. Ginzburg. 2010. Analogical thinking in ecology. *Quarterly Review of Biology* 85(2): 171–182.

Egler, F.E. 1986. Physics envy in ecology. *Bulletin of the Ecological Society of America* 67(3): 233–235.

Frege, G. 1970. In *Translations from the philosophical writings of gottlob frege*, ed. P. Geach and M. Black. Cambridge: Blackwell.

Ginzburg, L., and M. Colyvan. 2004. *Ecological orbits: How planets move and populations grow*. Oxford: Oxford University Press.

Gotelli, N.J. 2001. *A primer of ecology*, 3rd ed. Sunderland: Sinauer Associates.

[16]I am grateful for the assistance of Alan Baker, Bob Batterman, Jim Brown, John Damuth, Lev Ginzburg, Paul Griffiths, Stefan Linquist, Aidan Lyon, and Tim Raez. Thanks also to the participants in the conference "Unreasonable Effectiveness? Historical Origins and Philosophical Problems for Applied Mathematics", held at All Souls College, Oxford in December 2008 and participants in the workshop "Causation, Dispositions and Probabilities in Physics and Biology" held at the University of Lausanne, Lausanne, Switzerland. Work on this paper was jointly supported by an Australian Research Council Future Fellowship (grant number: FT110100909) and by a visiting fellowship to the Stellenbosch Institute of Advanced Study.

Grattan-Guinness, I. 2008. Solving Wigner's mystery: The reasonable (though perhaps limited) effectiveness of mathematics in the natural sciences. *The Mathematical Intelligencer* 30(3): 7–17.

Hales, T.C. 2001. The honeycomb conjecture. *Discrete Computational Geometry* 25: 1–22.

Hamming, R.W. 1980. The unreasonable effectiveness of mathematics. *American Mathematics Monthly* 87: 81–90.

Inchausti, P., and L.R. Ginzburg. 2009. Maternal effects mechanism of population cycling: A formidable competitor to the traditional predator-prey view. *Philosophical Transactions of the Royal Society B* 364: 1117–1124.

Israel, G. 1988. On the contribution of Volterra and Lotka to the development of modern biomathematics. *History and Philosophy of the Life Sciences* 10: 37–49.

Kingsland, S.E. 1985. *Modelling nature: Episodes in the history of population ecology*. Chicago: University of Chicago Press.

Levins, R. 1966. The strategy of model building in population biology. *American Scientist* 54: 421–431.

Lyon, A., and M. Colyvan. 2008. The explanatory power of phase spaces. *Philosophia Mathematica* 16(3): 227–243.

May, R.M. 2004. Uses and abuses of mathematics in biology. *Science* 303(6 February): 790–793.

Mirowski, P. 1989. *More heat than light: Economics as social physics, physics as nature's economics*. Cambridge: Cambridge University Press.

Ruben, D.H. 1990. *Explaining explanation*. London: Routledge.

Smart, J.J.C. 1990. Explanation-opening address. In *Explanation and its limits*, ed. D. Knowles. Cambridge: Cambridge University Press.

Sober, E. 1983. Equilibrium explanation. *Philosophical Studies* 43: 201–210.

Steiner, M. 1998. *The applicability of mathematics as a philosophical problem*. Cambridge: Harvard University Press.

Wigner, E.P. 1960. The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics* 13: 1–14.

Wilson, M. 2000. The unreasonable uncooperativeness of mathematics in the natural sciences. *Monist* 83(2): 296–394.

Comment on "The Undeniable Effectiveness of Mathematics in the Special Sciences"

Tim Räz

1 Introduction

The philosophical debate on the applicability of mathematics in the sciences goes back to the famous paper "The unreasonable effectiveness of mathematics in the natural sciences" by Eugene Wigner. Wigner felt a sense of miracle when thinking about the use of mathematics in physics. However, to this day, the use of mathematical methods in the *special* sciences has not been commonly accepted as a "wonderful gift", but has been challenged, if not outright rejected. This is the point of departure of Mark Colyvan's paper. His goal is to establish that mathematical models in the special sciences are not only helpful for prediction, but that they can play a genuinely explanatory role. This claim is supported by various case studies from biology, and especially population ecology.

I agree with Colyvan's main thesis that mathematical models in general, and models in population ecology in particular, can be explanatory, and that there are good reasons to accept the use of mathematical models in the special sciences. I will argue in Sect. 3 that Colyvan's main thesis can be supported historically – explanatory concerns played a key role in the genesis of mathematical population ecology. The historical fathers of mathematical population ecology, Vito Volterra and Umberto d'Ancona, anticipated Colyvan's thesis.

However, I disagree with Colyvan's account of some of the case studies. First, the issue of idealization, one of the main problems of (mathematical) modeling, is not sufficiently emphasized. I flesh this out with systematic considerations and by

T. Räz (✉)
Department of Philosophy, University of Lausanne, Quartier UNIL-Dorigny, Bâtiment Anthropole, 1015 Lausanne, Switzerland
e-mail: Tim.Raz@unil.ch

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_6,

consulting the historical sources. Volterra's discussions of the predator-prey model shows that he was acutely aware of the problem of idealization. I discuss this point in Sect. 4. Secondly, I argue in Sect. 5 that one of the cases is scientifically inadequate: the purported example of a mathematical explanation of the structure of the bee's honeycomb using the mathematical honeycomb conjecture is flawed.

2 Colyvan's Program

In the beginning, Colyvan points out that there is not *one* problem of applicability of mathematics to empirical science, but rather a multitude of different, but related problems. Accordingly, he restricts attention to one particular aspect of applicability, mathematical modeling in the special sciences in general, and population ecology in particular.

Colyvan's main problem arises from the tension between two facts. On the one hand, mathematical models in population ecology leave out, or abstract from, causal details, yet, on the other hand, they succeed in explaining biological phenomena. The task that Colyvan sets himself is to reconcile these two facts.

His focus is not on predictive models that merely reproduce phenomena but on explanatory models. The thought here might be that a model cannot be explanatory unless it somehow mirrors its target system. The puzzle, then, can be rephrased as: how can a model explain biological phenomena if it does not take into account biological facts?

If we want this to be a real puzzle, we have to presuppose a certain liberalism about explanations: we should not limit explanations to being causal, as this would render explanations that draw heavily on mathematics utterly mysterious. There are good reasons for not making any such assumption; most importantly, there appear to be genuine examples of scientific explanations that are non-causal.

Colyvan details three ways in which mathematical models in the special sciences can be explanatory. First, mathematical models do not leave out all details, but sum up and represent some aspects of the modeled systems in mathematical form. Second, abstracting away from (causal) details can be explanatorily advantageous; abstraction can help in drawing analogies and thereby facilitate inferences. The third, and most contentious, claim is that in some examples, mathematics carries the bulk of the explanatory burden; according to Colyvan, such explanations deserve to be called mathematical explanations.

These claims are supported with some well-known case studies. In the remainder of this paper, I will comment on two of these case studies, the predator-prey model and the honeycomb case.

3 The Roots of Lotka-Volterra

The predator-prey model is Colyvan's main example of a mathematical model from population ecology, see Colyvan, Sect. 2 of chapter "The Undeniable Effectiveness of Mathematics in the Special Sciences", this volume. The model consists of a system of two coupled differential equations that describe how predator and prey populations evolve over time if we view them as an isolated system. It is a basic, yet powerful model that serves as a template for more complex models.

In this section, I revisit Vito Volterra's and Umberto d'Ancona's original publications on population ecology, with special attention to the predator-prey model. As they tell the story, what prompted the construction of the model was a request for an explanation. Also, Volterra and d'Ancona had a clear methodological motivation for choosing the path of mathematical modeling. They explicitly address the relation between modeling in biology and physics and "physics envy".

Volterra first proposed the predator-prey model to solve a puzzle that Umberto d'Ancona, a marine biologist and Volterra's son-in-law, brought to his attention.[1] Fishery statistics showed an increase in the number of predators relative to the number of prey in the adriatic sea during the first world war, a period of diminished fishery; the previous proportion was restored when fishery was back to pre-war intensity.

Volterra was able to qualitatively reproduce and explain this surprising fact with the predator-prey model. He first inferred from the model that the population sizes oscillate indefinitely around a fixed equilibrium point if the system is undisturbed. He then established that the equilibrium point is the time average of the population sizes. Finally, he examined how the equilibrium point is affected by a general, continuous biocide, such as fishery. He found that on average, such a biocide increases the number of prey and decreases the number of predators. This result, which is known as Volterra's third law, explains the fishery statistics: if the general biocide is suspended, the average of the predator population increases, while the average of the prey population decreases.

The lesson we can learn here is that the predator-prey model was not constructed merely to reproduce a certain phenomenon, but to explain a high-level feature of a system of fish populations. The project of mathematical modeling in population ecology was explanatory from the start. What is more, the mathematical model only gives us an approximate, qualitative understanding of the population interactions, as the coupled, nonlinear differential equations generally cannot be solved exactly. The material motivation for the model, as well as its mathematical features, lead to a qualitative understanding of the system under scrutiny – precise quantitative predictions are neither the motivation behind the model, nor are they possible.

[1] See Volterra (1928) for a historical exposition of the model; the original motivation is mentioned on p. 4, footnote 2.

Second, let us have a closer look at the methodological motivations driving Volterra's and d'Ancona's research.[2] In Volterra and D'Ancona (1935, Chap. 1), the two authors reflect on the reasons for adopting modeling as their method of choice. In a nutshell, Volterra and d'Ancona would have preferred a different, more direct approach to population ecology, but were forced to adopt the modeling path by limited epistemic access to the system under scrutiny.

Their first preferred method would have been controlled experimentation. This, however, is not feasible because actual populations cannot be controlled adequately: they are spread out spatially, their breeding cycles are too long, and environmental factors vary indefinitely. All this prevents a direct, experimental approach. Second, detailed statistics could compensate for some of the epistemic limitations of controlled experimentation, as varying factors would eventually cancel out over time. However, a statistical approach is out of the question too, because the necessary means to carry out statistical evaluations were not available at the time.

As the methods of experimental control and statistics cannot be applied, Volterra and d'Ancona propose a third, more indirect approach. They write:

> Since it appears too difficult to carry through quantitative studies by experiments and thus to obtain the laws that regulate interspecific relationships, one could try to discover these same laws by means of deduction, and to see afterwards whether they entail results that are applicable to the cases presented by observation or experiment.[3]

Volterra and d'Ancona advocate an indirect, modeling approach. However, they only resort to mathematical modeling *faute de mieux*: a more direct approach based on controlled experiments is simply too difficult to carry out. The use of mathematics is not due to the desire to apply complicated mathematics at all cost, but due to a lack of alternatives. This means that in order to carry out their explanatory project, the use of modeling techniques was the only viable option. Population ecology faced real methodological difficulties that could only be overcome with the help of mathematical methods.

These considerations also partially answer the question whether Volterra and d'Ancona should be accused of physics envy, "the mistake of using inappropriate – and, in particular, overly-complicated, inappropriate, and physics-inspired mathematics – in ecology" (Colyvan, Sect. 4 of chapter "The Undeniable Effectiveness of Mathematics in the Special Sciences", this volume). In view of their methodological reflections, Volterra's and d'Ancona's use of mathematical modeling is not due to the sheer intellectual pleasure of using complicated mathematical methods, it

[2]The following discussion of Volterra's and d'Ancona's methodological reflections is based on Scholl and Räz (2013, Sect. 3).

[3] D'ailleurs s'il apparaît trop difficile d'effectuer l'étude quantitative par voie d'expérience et d'obtenir ainsi les lois qui règlent les rapports interspécifiques dans les associations biologiques, on pourra tenter de découvrir ces mêmes lois par voie déductive et de voir ensuite si elles comportent des résultats applicables aux cas que présente l'observation ou l'expérience. (Volterra and D'Ancona 1935, p. 8)

is an attempt to overcome practical problems of other, more direct approaches to population ecology.

Volterra and d'Ancona are outspoken defendants of the use of mathematical methods in population ecology that parallel physics. After all, Volterra held a Chair of Mathematical Physics at the University of Rome, and he does not hide his sympathy for mathematical methods:

> One should not worry too much when one considers ideal elements and imagines ideal conditions that are not completely natural. This is a necessity, and it is sufficient to think of the applications of mathematics to mechanics and physics that have lead to results that are important and useful in practice. In rational mechanics and in mathematical physics one considers surfaces without friction, absolutely flexible and unextended strings, ideal gases, and so on. The example of these sciences is a great example we should always keep in mind and that we should strive after.[4]

Two aspects of this passage stand out. Volterra and d'Ancona recommend the example of mathematical physics and its successes as a template for other sciences, especially biology. However, this is not an unqualified endorsement, as we have seen above. There are reasons to adopt mathematical methods beyond the fact that this has worked well in the case of classical mechanics. It would be unfair to accuse them of physics envy – even more so as they are acutely aware of the dangers and pitfalls of modeling.

Secondly, Volterra and d'Ancona point out that "ideal elements and conditions" are an integral part of physics, and the same is true in biology. Their emphasis on the issue of idealization should be taken seriously, as they discuss the issue not only in the methodological reflections, but bring it up over and over again.

4 Volterra and d'Ancona on Idealization

Colyvan comments on idealization[5] in his first answer to the objection that mathematical models fail to give an adequate account of ecological systems, see Colyvan, Sect. 2 of chapter "The Undeniable Effectiveness of Mathematics in the Special Sciences", this volume. He distinguishes between the claim that mathematical

[4] D'autre part, il ne faut pas trop se préoccuper si on envisage des éléments idéaux et l'on se place dans des conditions idéales qui ne sont pas tout à fait ni les éléments ni les conditions naturelles. C'est une nécessité et il suffit de rappeler les applications des mathématiques à la mécanique et à la physique qui ont amené à des résultats si importants et si utiles même pratiquement. Dans la mécanique rationnelle et dans la physique mathématique on envisage en effet les surfaces sans frottement, les fils absolument flexibles et inextensibles, les gaz parfaits, etc. L'exemple de ces sciences est un grand exemple que nous devons avoir toujours présent à l'esprit et que nous devons tâcher de suivre. (Volterra and D'Ancona 1935, p. 8)

[5] The debate on idealizing models goes at least back to Cartwright (1983). A useful overview of idealization in the context of modeling can be found in Frigg and Hartmann (2012, Sects. 1.1 and 5.1).

models fail to represent at all, and the claim that mathematical models are overly simple, or misrepresent. He refutes the first claim by pointing out that the parameters of the predator-prey model can be interpreted as summing up biological information such as birth and death rates.

In his response to the second claim, he grants that the original predator-prey model is overly simple, but points out that its role in modern population ecology is pedagogical, and that it can serve as a template for more sophisticated models. He thinks that even these models leave out biological details, which, however, does not invalidate them, as leaving out details is part and parcel of the practice of modeling. In sum, Colyvan considers the first claim to be true, but somewhat beside the point.

I agree with Colyvan that the predator-prey model does sum up and represent some biological information. The model's parameters have a clear biological correlate. However, I wonder whether anyone has ever actually defended the claim that the predator-prey model fails to represent completely. I find the claim that (some) mathematical models are overly simple and misrepresent to be much more interesting, and troubling.

Colyvan's answer to the misrepresentation charge is that the use of abstract models is an integral part of modeling, if not of science. I think that Colyvan's answer is correct, but it does not address the issue of idealization. Idealizing models that misrepresent their target system should be contrasted with models that merely abstract from some aspects of their target system, or leave out certain properties: Idealizing models violate a "nothing but the truth" clause, while models that abstract violate a "the whole truth" clause.[6] To emphasize the legitimacy of abstract models will not save the defenders of mathematical models when accused of idealization. The real challenge is to tell a story about how models that lie about their target system can nevertheless be explanatory.

The claim that the issue of idealization is much more pressing than abstraction is supported by a close reading of Volterra's and d'Ancona's original publications. We already saw in the last quote above that Volterra and d'Ancona were aware of the fact that they made ample use of "ideal elements and conditions". They defend this practice by pointing out that idealizations are very common in physics. However, we do not have to rely on these general remarks: The use of idealizations is acknowledged and defended throughout Volterra (1928) and Volterra and D'Ancona (1935). Here are two examples.

In Volterra and D'Ancona (1935, Chap. 3), the principles that form the basis of a mathematical approach to population ecology are introduced. The authors justify the use of continuous variable models as follows:

> There is no need to remind ourselves that in reality, the number of individuals of the species living together varies in a discontinuous manner and always in integer numbers. But on a mathematical approach it is convenient to assume that the variation is continuous in order to

[6] See e.g. Batterman (2010) for a recent discussion of the distinction between abstraction and idealization.

> be able to apply the methods of infinitesimal calculus. That is why the number of individuals is not an entire number, but any real, positive number which can vary continuously.[7]

Volterra and d'Ancona are aware of the fact that the use of continuous variables is an idealization. The use of this idealization is justified pragmatically – differential equations are very well understood and therefore more convenient in application than a discrete approach.

Later, they discuss an even more substantial idealization. The predator-prey model uses constant growth coefficients to describe the part of the dynamics that is independent of interactions (these are the linear parts of the predator-prey equations with coefficients r and q, see Colyvan, Sect. 2 of chapter "The Undeniable Effectiveness of Mathematics in the Special Sciences", this volume). Volterra and d'Ancona discuss the use of constant growth coefficients earlier in the text:

> To simplify the problem, we assume constant [growth] coefficients, even though we know that in reality they never are constant, as the coefficient of birth, as well as the coefficient of death, varies with the age of the individual. We thus assume that the number of births and of deaths is proportional to the total number of individuals living at a given moment.[8]

Here it sounds as if the only reason to use constant coefficients is to make things easier for the modeler. Later on, however, Volterra and d'Ancona attempt to justify this simplifying assumption: constant growth coefficients may be harmless "within certain limits", but the limitation always has to be kept in mind, as it can lead to an infinite growth of the prey population.

Continuous variables and constant growth coefficients are only two examples where Volterra and d'Ancona address the use of idealizations in the predator-prey model. They are acutely aware of the problem and try to justify it by either granting that their choices are pragmatic, or by arguing that the idealizations are to a certain degree harmless. The fact that they continually bring up the issue of idealization goes a long way to show that the question of misrepresentation is much more pressing than the question whether models represent at all. Idealization is one of the problems Volterra and D'Ancona took seriously.

---

[7] Il est inutile de rappeler qu'en réalité le nombre des individus formant les espèces vivant ensemble varie d'une manière discontinue et toujours par nombres entiers. Mais dans l'étude mathématique il convient de supposer des variations continues afin de pouvoir appliquer les procédés du calcul infinitésimal; c'est pourquoi le nombre des individus est considéré non pas comme un nombre entier, mais comme un nombre réel et positif quelconque et variant par degrés continus. (Volterra and D'Ancona 1935, p. 14)

[8] Pour simplifier le problème on suppose constants ces deux coefficients, alors qu'on sait bien qu'en réalité ils ne le sont jamais, puisque le coefficient de natalité, aussi bien que celui de mortalité, varient avec l'Age de l'individu. On suppose donc que le nombre des naissances et celui des décès sont proportionnels au nombre total des individus en vie à tel moment donné. (Volterra and D'Ancona 1935, p. 16)

5 A Legitimate Explanation of the Honeycomb?

Colyvan adduces the explanation based on the honeycomb conjecture to support the claim that there are mathematical explanations of physical phenomena, that is, explanations where the mathematics does most of the explanatory work.[9] I do not question this claim as such, but merely argue that this particular explanation is not scientifically adequate, and that it therefore does not support Colyvan's claim.[10]

The idea behind the explanation is the following. The mathematical honeycomb conjecture states that a hexagonal grid is the most efficient way to tile a two-dimensional surface into equal regions. This, together with the biological premiss that it is an evolutionary advantage for the honeybees to minimize the use of wax in the construction of honeycomb cells, explains the fact that the bee's honeycomb has its structure.

It is probably an evolutionary advantage for the bees to minimize the use of wax, so some form of (mathematical) optimization may well be relevant here. The optimization problem has to take into account that the bees need access to their cells; therefore, as Colyvan points out, a general, thee-dimensional sphere-packing problem, or cell-packing problem, is not the right mathematical formulation of the problem.

The problem with the explanation based on the honeycomb conjecture is that it applies to two-dimensional surfaces and therefore can only take the shape of the openings of the cells into account. This, however, is not sufficient. It is not clear that a structure with optimally shaped openings minimizes the amount of wax. A structure can have cells with optimal openings, but still have some wild shape otherwise. We cannot infer the optimality of cells from the shape of the openings. To make sure that a structure is optimal, we have to take it into account as a whole.

The relevant optimization problem, then, is three-dimensional. What is minimized is the amount of wax relative to cells of unit volume. Additionally, the optimization problem has to satisfy certain boundary conditions. One of these is that each cell needs an opening of reasonable size. A possible mathematical formulation of the problem is as a bounded form of the Kelvin problem, the optimal tiling of space with cells of equal volume, with the restriction that the cells lie between two parallel planes such that each cell has an opening in one of the planes.[11] This sort of optimization could be relevant to the actual honeycomb.

However, the trouble with this kind of three-dimensional optimization problem is that they have not been solved – mathematics just is not there yet – and as we

[9] The explanation was first proposed in Lyon and Colyvan (2008).

[10] This section draws on Räz (2013). A detailed account of the objections against the adequacy of the explanation can be found there.

[11] This kind of problem was proposed and analyzed in Fejes Tóth (1964).

do not know the solutions, we do also not know whether the optimal structure has hexagonal openings.

Summing up, there may well be a mathematical explanation of the structure of the bee's honeycomb involving some optimal geometrical structure. However, it cannot be (exclusively) based on the two-dimensional honeycomb conjecture. Also, it is not clear that the hexagonal tiling is part of the relevant optimal three-dimensional structure.

6 Conclusion

Colyvan's claim that mathematical models in the special sciences have an important, and even explanatory role, is supported by historical facts; population ecology relied on mathematical models for explanatory purposes from the beginning. The fathers of population ecology, Volterra and d'Ancona, also defended themselves against the charge of physics envy; at least, they were aware of the problem.

Some details of Colyvan's case studies are problematic. I tried to show that one of the most important problems with mathematical models is idealization; I think that this should be emphasized more. Then, I argued that the honeycomb explanation is flawed. This is a mistake that has sneaked into the philosophical debate, and we should stop using the case as an example of a mathematical explanation. However, this does not invalidate Colyvan's general philosophical claims.

The debate on the applicability of mathematics raises issues that deserve attention from both philosophers of science and scientists. As Colyvan points out, there is not one unified problem of applicability, but many related questions. The most fruitful approach to these questions is probably via more detailed and historically informed case studies.

Acknowledgements I thank Mark Colyvan, Michael Esfeld and Raphael Scholl for useful comments on earlier drafts of this paper. Parts of Sects. 3 and 4 are based on joint work with Raphael Scholl and the paper (Scholl and Räz 2013). This work was supported by the Swiss National Science Foundation, [100018-140201/1].

References

Batterman, R.W. 2010. On the explanatory role of mathematics in empirical science. *The British Journal for the Philosophy of Science* 61: 1–25.

Cartwright, N. 1983. *How the laws of physics lie*. Oxford: Oxford University Press.

Fejes Tóth, L. 1964. What the bees know and what they do not know. *Bulletin AMS* 70: 468–481.

Frigg, R., and S. Hartmann. 2012. *Models in science*. http://plato.stanford.edu/entries/models-science/.

Lyon, A., and M. Colyvan. 2008. The explanatory power of phase spaces. *Philosophia Mathematica* 16(2): 227–243.

Räz, T. 2013. On the application of the honeycomb conjecture to the bee's honeycomb. *Philosophia Mathematica* 21: 351–360.

Scholl, R., and T. Räz. 2013. Modeling causal structures. *European Journal for Philosophy of Science* 3(1): 115–132.

Volterra, V. 1928. Variations and fluctuations of the number of individuals in animal species living together. *Journal du Conseil International pour l'Exploration de la Mer* 3(1): 3–51.

Volterra, V., and U. D'Ancona. 1935. *Les associations biologiques au point de vue mathématique*. Paris: Hermann.

Part II
Philosophy of the Natural and Life Sciences

Explanatory Pluralism in Psychiatry: What Are We Pluralists About, and Why?

Raffaella Campaner

1 Models of Psychiatric Disorders. Some Studies from Psychiatry

How psychiatric disorders can be defined, classified and understood are extremely problematic issues and the object of growing debate taking place in the philosophy of psychiatry and involving the philosophy of science more in general. A number of themes and perspectives from different fields intertwine in this debate: neurology, neuroscience, genetics, ethics, psychology, cognitive science, and others participate from different standpoints in an effort to shed light on the status and etiology of mental pathologies.[1] Investigations are carried on with different theoretical frameworks and focuses, which are rendered more or less explicit. A glimpse at the scientific literature shows that research on psychiatric disorders is plentiful and diverse: various kinds of evidence are collected and evaluated; different levels and interacting variables are investigated; different disciplinary fields are made to cooperate; the focus can be placed on populations, subclasses of populations, or individuals.

The "dappled nature of causes of psychiatric illness" (Kendler 2012a) inspires a variety of conceptions, leading in turn to different ways of dealing with the patient. Studies in the last few decades have been performed from within psychiatry on different models of pathologies as actually employed in medical practice. I will start

[1] I will not address here the long-standing and ongoing debate on the different versions of the "Diagnostic and Statistical Manual of Mental Disorders" (DSM). On this see e.g. Kendler and Parnas (2012).

R. Campaner (✉)
Department of Philosophy and Communication, University of Bologna, Via Zamboni 38, 40126 Bologna, Italy
e-mail: raffaella.campaner@unibo.it

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_7,

by briefly presenting a few examples showing – amongst others – that a crucial role in the adoption of a given model is played by the context in which it is employed and the function it is meant to perform.

In a recent study on psychiatrists' concepts of mental illness and their underlying attitudes (Harland et al. 2009), a group of trainee psychiatrists were asked to answer a questionnaire on four psychiatric disorders: schizophrenia, antisocial personality disorder, generalized anxiety disorder and major depressive disorder. Their attitudes toward eight different approaches to mental disorders, as referred to the four pathologies, were assessed. The approaches can be sketchily summarized as:

- *Biological*: the disorder results from brain dysfunction and underlying biological abnormalities;
- *Cognitive*: the disorder is the sum of maladaptive thoughts, beliefs and behaviors;
- *Behavioral*: the disorder results from maladaptive associative learning;
- *Psychodynamic*: the disorder results from the failure to successfully complete developmental psychic stages;
- *Social realist*: the main causes of the disorder are social factors such as prejudice, poor housing and unemployment;
- *Social constructionist*: the disorder is a culturally determined construction that reflects the interests and ideology of the socially dominant groups;
- *Nihilist*: attempts to explain the disorder in rigorously scientific terms have obtained no significant knowledge;
- *Spiritual*: the disorder is due to the neglect of the spiritual or moral dimensions of life.

Although responses varied, overall the study indicates the biological model as the most strongly endorsed, with schizophrenia being the disorder about which convictions were expressed most forcefully. However, it interestingly emerges that model endorsement varies with the disorder considered, suggesting that not all psychiatric pathologies are etiologically considered on a par. For instance, schizophrenia is mainly believed to have a biological etiology and hence to need to be investigated through biological research, whereas generalized anxiety disorder is largely regarded as resulting from maladaptive thoughts and beliefs. Moreover, while the trainees turn out to be more committed to the biological model, they do not emerge as exclusively committed to any one model, revealing that a multiplicity of views can be supported within the same group, with the same scientific training: "as a group, they organize their attitudes towards mental illness in terms of a biological/non-biological contrast, an 'eclectic' view and a psychodynamic/sociological contrast" (Harland et al. 2009, p. 967).

A previous study with analogous results was presented by Michael Brog and Karen Guskin (1998), who focused on medical students' view of etiology and the treatment of disorders. Their study claimed to show that – contrary to the authors' own expectations – third-year medical students are able to take into account both biological and psychological factors when dealing with psychiatric illness, hence tending to opt for a combined approach of medication and psychotherapy

in treatment. If this holds in general, though, medical students considered in the study did not regarded *all* psychiatric disorders as lying at the same place along the biological-psychological spectrum: e.g., schizophrenia and bipolar affective disorder were seen as significantly more biological in nature, whereas narcissistic and borderline personality disorders were seen as significantly more psychological in nature.

The adoption of different etiological models of disorders can have many practical implications, affecting, among others, management and decision-making in multi-agency teams. That different health professionals and service users embrace different views – which is then very likely to result in difficult communication – has been argued, e.g., in Colombo et al. (2003). The study here illustrated was performed on 100 participants representing psychiatrists, community psychiatric nurses, approved social workers, patients and informal carers operating in Leicestershire (UK). Six models of mental disorder were identified, and each was defined on various dimensions by asking:

(a) what the nature of a mental disorder is (how it can be defined and diagnosed; how the behaviour can be interpreted; how it can be labeled; what its etiology is);
(b) what should be done about it (what treatment should be adopted; what function a psychiatric hospital has; what prognosis can be elaborated);
(c) how people involved (practitioners, informal carers, society, patients) should behave towards each other (what their rights and duties are).

Here too attitudes towards schizophrenia were compared, showing that "each of the study's multi-agency groups *implicitly* supports a complex range of model dimensions regarding the nature of schizophrenia, the appropriateness of specific forms of treatment and care, and their respective rights and obligations towards each other" (Colombo et al. 2003, p. 1557, italics added). Individuals answered differently, with people also within the same professional group embracing different models. However, overall psychiatrists and community psychiatric nurses favored the medical approach, social workers implicitly strongly endorsed the social model, and the group of patients exhibited large disparity in the perspectives embraced, thus probably hinting at a greater heterogeneity within such a group. All this is very relevant with respect to the power relationships holding between various practitioner groups, as well as between practitioners and patients in clinical encounters. I am not claiming that diagnoses and treatments require a precise definition and full-fledged explanatory theories of diseases, but that the underlying accounts – whatever their level of detail – have an impact on both research trends and clinical practice. It is very important to make implicit models explicit, given the tangible implications that their adoption can have in dealing with the disease and treating it: models cannot but strongly inform decisions, therapies and, where possible, prevention strategies. It is recognized that the better the communication and the more consistent the purposes of mental health and social service agents involved in a clinical context, the more successful their activity is. Despite the importance of these aspects for

the welfare of patients and public health, models of mental disorders – often only implicitly embraced – continue to differ significantly. Differences also hold between medical statements in textbooks and practitioners' assumptions, between medical and lay conceptualization of disease, between processing of information by practitioners and by patients. "In clinical practice, implicit adherence to a model may interfere with team decision making by generating conflicting assumptions that create misunderstandings between multi-agency groups" (*id*., p. 1558). Colombo et al. suggest that the consistency of purposes, the procedures promoted and service delivered could be enhanced by greater sharing of the models adopted.

Another study on models actually employed by practitioners in psychiatry has analyzed self-report questionnaires (answered by 127 out of 270 psychiatrists and psychologists in the Department of Psychiatry at McGill University, Quebec, Canada), and considered both the attitudes towards mental diseases' etiology and the levels of intentionality, controllability, responsibility and blame attributed to the patients. The pathological conditions considered were: (1) a manic episode induced by selective serotonin reuptake inhibitors (seen as a biologically determined process); (2) narcissistic personality disorder (seen as psychological); (3) heroin dependence (seen as falling somewhere in between the biological and the psychological realm). Data were taken to reflect how psychiatrists often continue to operate according to a dualistic mind-brain perspective in ways that may be covert and unacknowledged. The adoption of a more genetics- and biology-oriented model, or, vice versa, of a model prone to stress the causal relevance of social and psychological factors can be influenced by the persistence of the dichotomy between mind and brain, with a noticeable impact on attributions of responsibility. Collected answers have led to the conclusion that "the more a behavioral problem is seen as originating in 'psychological' processes, the more the patient tends to be viewed as responsible and blameworthy for his or her symptoms; conversely, the more behaviors are attributed to neurobiological causes, the less likely patients are to be viewed as responsible and blameworthy" (Miresco and Kirmayer 2006, p. 913). Clinicians involved in the study tended to associate psychological causation of mental illness with intentionality, controllability and responsibility, and hence to blame the patients for their anomalous actions, and biological etiology with unintentional and uncontrollable behaviour, which was then regarded to fall outside the patients' sphere of personal responsibility.

All these studies highlight how the scenario in which psychiatric practitioners act and patients are treated is far from homogeneous; despite much debate, etiological levels are often kept separate, with explanations focusing on specific levels, and some oppositions still hold. "Reductive explanatory models can emerge from either end of the mind-brain spectrum. [. . .] Patients may be short-changed by hard-nose and inflexible diagnostic and therapeutic approaches that are reductionist in either the biological or the psychological direction" (Brendel 2003, p. 567). Contrasts can influence trends of development of scientific theories in the field, and affect the choice of the clinical management of the disorder.

In the scenario described, plurality appears in a number of respects:

- a plurality of disciplinary fields are involved;
- a plurality of kinds of evidence is collected to support different models (e.g. biomolecular research; epidemiological studies; first-person reports, ...)[2];
- a plurality of focuses are present (focus on populations, subclasses of populations, individuals);
- a plurality of explanatory models, concentrating on different etiological levels, are employed
  - for different disorders,
  - by a plurality of figures involved - as agents and patients,
  - with a plurality of aims (research, treatment, prevention, attribution of responsibility, ...).

Current plurality of approaches, its origin and meaning, can be interpreted in different ways, in itself and as a general feature of psychiatry as a discipline.

> It may be that even where consensus seemingly exists (for example, that schizophrenia is best understood through the biological model), this may itself represent a cultural lag between the attitudes of clinicians (including trainees) and the evidence base of current as yet unknown research, which, for example, suggests schizophrenia to be a complex multifactorial disorder with important environmental and social constraints on etiology. [...] An alternative interpretation is that one of psychiatry's great strengths is that it draws freely on different intellectual disciplines and should therefore be viewed as a 'multi-paradigm' science (Harland et al. 2009, p. 976).

In the next section we shall see how the debate on these issues intertwines with some recent reflections on explanatory processes as developed within psychiatry, and interestingly intersects with the current philosophical debate on the topic.

2 Approaches to Explanation of Psychiatric Disorders

Reflections on general approaches to explanation employed to uncover the etiology of psychiatric disorders have been expanding significantly. Far from exhausting all the views discussed at present, this section reviews some recent proposals elaborated by some of the most authoritative authors in the field, such as Dominic Murphy, Kenneth Kendler and John Campbell.

According to Dominic Murphy (2011), the *medical model* is the most widespread view on psychiatric disorders. In very general terms, this model sees mental illnesses as brought about by pathological neuropsychological processes, and can come into different versions. The *minimal interpretation* regards diseases as collections of symptoms occurring together and unfolding in characteristic ways, without making any commitment with respect to underlying causes. According to the *strong*

[2] On this see e.g. O'Connor et al. (2012).

interpretation of the medical model, mental illnesses are brought about by specific pathophysiological processes in the brain, and causal explanations indicate how such processes take place and how they give rise to the clinically observable symptoms of mental illness. Instead of restricting attention to symptoms, as the minimal interpretation does, the strong interpretation searches for underlying causal pathways and conditions. The same symptoms may be produced by different underlying causes and conditions and, vice versa, the same causal pathways can result in different individual surface features. Progress in understanding the brain's physiology will result in improvements in clinical practice too.[3] The strong interpretation responds to the idea that mental disorders are not to be dealt with simply in terms of conceptual varying constructs, but as genuine biological entities, with distinctive pathophysiological features.[4]

Murphy endorses the medical model, regarding it as the most adequate to the target of making psychiatry a successful and fully established science. He takes as his background theory cognitive neuroscience, very broadly conceived as a general framework to understand mental life in terms of information processing systems in the nervous system, investigated via a number of theories at different levels. Contrary to the idea that the medical model per se privileges some level, Murphy maintains that, if properly construed, it can include variables belonging to as many levels as necessary, without any commitment whatsoever to reductive explanations in terms of molecular biology.[5] Murphy hence stresses how pathogenic neurobiological processes are necessary but not sufficient for an understanding of mental illness and argues against a geneticisation of psychiatry. The role of genes in the development of mental disorders is very controversial; their impact on individual risk has yet to be explored, is probably rather small and often nonspecific as well as highly dependent on environmental exposure. An example is given by research on major depression, which emphasizes how both particularly stressful life events – e.g. humiliation (see Kendler and Prescott 2006) – and genetic factors are depressogenic. "One might think that major depression just is some, as yet unknown, cognitive and/or neurological process (or perhaps a family of specific processes) that can be triggered in diverse ways depending on one's genetic inheritance, acquired psychology and contingent biography" (Murphy 2011, p. 436).

Murphy suggests that in psychiatry we use *models* to explain *exemplars*, which are idealized representations of the symptoms of disorders and the course they have. Exemplars take collections of symptoms to unfold over time in analogous

[3] See e.g. Andreasen (1997) and Black (2005), cited in Murphy (2011).

[4] Murphy relates the divide between the strong and the minimalist views to the divide over realism about disease: "we do better, if we are to be realists about disease categories, to view diseases as realized in biological systems [. . .], and this permits strong medical thinking to acknowledge that a realization which is shared across patients might have a variety of specific, peculiar causes" (Murphy 2011, p. 436).

[5] Others, e.g. Patil and Giordano (2010), maintain instead that reductionism, granting genetic and biochemical entities explanatory primacy, is one of the core ontological assumptions of the medical model in psychiatry.

ways, and take patients to respond similarly to the same treatments. They orientate inquiry and, insofar as they are idealizations of disorders and in this sense represent imaginary patients, abstract from individual variations. The construction of exemplars is meant to set "the causal-explanatory challenge". Models are built to represent the pathogenic process accounting for the observed phenomena in the exemplar, where many symptoms jostle together at different levels of explanation.[6] Exemplars are narratives; causal relations underlying them are claimed to do the explaining. Observed relations are explained by identifying the mechanisms that bring them about. To explain, we appeal to our mechanistic knowledge concerning what are regarded as standard forms of behaviour. Mechanistic scientific theories accounting for the ways in which cognitive parts work and interact are employed to explain abnormal outcomes as resulting from the organism's failure to function normally. A general theory of standard neurological functioning of the brain is hence presupposed as a ground for explanations of psychiatric disorders. Clinical reasoning analyses to what an extent exemplars resemble real patients.

The mechanistic perspective has been popular in the medical context. "Traditionally, medicine has been successful in establishing etiology of diseases and disorders, and developing focal therapies based upon [. . .] *mechanistic conceptualizations*" (Patil and Giordano 2010, p. 1, italics added). If Murphy suggests that observed relations modelled in exemplars are to be explained by identifying underlying mechanisms, Kenneth Kendler (2008) maintains that the understanding of mechanisms should be seen as an appropriate scientific model for psychiatry insofar as a mechanistic approach is naturally suited to a multicausal framework. The model works by decomposing complicated mechanisms into simple subunits, allowing them to be studied in isolation and then to reassemble the constituent parts into their functioning wholes. While this operation can be rather straightforward when dealing with additive mechanisms, it is much more problematic in a field like psychiatry, where causal networks investigated present multiple nonlinear interactions and causal loops. Psychiatry has been struggling for a long time with the interrelationships between biological and psychological explanatory perspectives, aiming to shed light on interactions between biological, psychological and socioeconomic processes. What is really at stake is that psychiatry does not demand to clarify biological, psychological or socio-cultural processes as such, but unique processes arising from some peculiar intertwining of such different kinds of processes. In this respect, Kendler claims, a mechanistic account can provide a middle ground between hard reduction and hard emergence,[7] and allows us to understand how pathologies are actually instantiated in the world. Decomposition is held to be driven by a reductionist stance, while rearrangement of constituent parts and their activities into complex wholes is guided by some sense of high-level

[6]For instance, symptoms of major depression present in the exemplar might include lowered affect serotonin imbalances, negative self-assessment, disturbed sleep, lethargy, lack of motivation.

[7]Kendler is thinking of a mechanistic approach such as Bechtel's (2007).

organization. Alcohol dependence provides a significant example. Its risk factors belong to at least four broad levels:

1. *biological/genetic*: liability to alcohol dependence may be influenced by prenatal exposure to alcohol and by aggregate genetic factors[8];
2. *psychological*: several personal traits, such as neuroticism, impulsivity and extraversion, can affect the risk for alcohol dependence;
3. *social*: social factors (e.g. drug availability) and social class can play a causal role;
4. *cultural/economic*: risk for alcohol dependence is affected by factors such as forms of alcohol commonly consumed in a social group, acceptability of public drunkenness, traditional cultural practices and religious beliefs, level of taxation of alcoholic beverages, statutes controlling sizes of alcoholic beverage containers, ...

Biological, psychological, and cultural factors do not affect risk independently, but impact on each other in various ways. For instance, "genes influence subjective ethanol effects, which influence alcohol expectations, which in turn loop out into the environment, influencing consumption patterns, which in turn affect risk of alcohol dependence" (Kendler 2008, p. 697). More generally, the actions of biological factors can be modified by the environment (e.g. light-dark cycle), stressful life experiences (e.g. maternal separation), and cultural forces (e.g. learning tasks). Hence, a hard reductive biological approach would not do and hopes are set in the implementation and integration of biological explanations with accounts articulated in the language of psychology. According to Kendler, a mechanistic explanatory approach, properly construed, does not confine explanatorily relevant factors to any single level (molecular or otherwise) and allows for – at least partial and local – decomposability. Decomposability could not only facilitate research, but also help meet clinical concerns, allowing for more focused and effective treatments and health service.

In roughly the same years, Kendler has suggested that a different general approach to scientific explanation, the interventionist approach, can provide an adequate explanatory framework. Together with John Campbell, he has highlighted the merits of the interventionist model in psychiatry, claiming, among others, that it:

– accomplishes the essential task of distinguishing between predictive-correlative and genuinely causal relationships;
– is non-reductive;
– is agnostic with respect to the mind-body problem, admitting of both relations from mind to brain and from brain to mind.

[8]See e.g. Molina et al. (2007), Dick and Bierut (2006), Edenberg et al. (2004), cited in Kendler (2008).

In an interventionist framework, conceptualizing causal factors as difference-makers allows them to be freely distributed across multiple levels (see Kendler 2012a), and the identification of causally relevant variables is separated from the elucidation of specific underlying mechanistic processes. At the same time, the idea of a mechanism is by no means to be seen as competing with the interventionist view, but, on the contrary, as supplementing it. What is insisted upon is that "the interventionist model can provide a single, clear empirical framework for the evaluation of all causal claims of relevance to psychiatry and presents psychiatry with a method of avoiding the sterile metaphysics arguments about mind and brain which have preoccupied our field but yielded little of practical benefit" (Kendler and Campbell 2009, p. 881). Not being involved with revealing the mechanisms underlying causal relations, the interventionist view is not loaded with commitments regarding the mind-brain issue. Furthermore, the most attractive features of the interventionist approach include its connecting causation with the practical interests of psychiatry, with treatment and prevention, by virtue of its defining causation in terms of what would happen under interventions. As also stressed in Campbell (2008), an interventionist view allows a pick and mix of variables at all levels (neural, genetic, psychological, economic, socio-cultural, ...), at the same time avoiding the risk of uncritically including causally irrelevant conditions. To evaluate the impact of interventions, background conditions – which may vary widely in psychiatric phenomena – will have to be made as explicit as possible. The manipulationist standpoint is also presented as broadly suitable for psychiatry, for instance, by Schaffner (2002), Woodward (2008), and Murphy (2011) himself. According to Murphy, the manipulationist view naturally fits medicine's aim to figure out what makes a difference in a given context, and its formal properties are perfectly compatible with the identification and interrelation of many sorts of variables, from the molecular to the cultural.

It is worth stressing that the motivations provided to support the different views sketched above coincide partly and address some of the most problematic aspects of psychiatric explanations. The different positions prospected by Murphy, Kendler and Campbell are strongly motivated by the aims of: including multiple explanatory levels, avoiding reduction to exclusively biological entities, and recognizing the irreducible explanatory role of high level variables such as, e.g., socio-economic variables. High-level explanations are held not to have a merely provisional and/or heuristic character, but to provide genuine and indispensable explanatory information. Furthermore, some abstracting from individual variations is inevitably required to provide general models of diseases to be employed in a number of specific instantiations, and properly adapted to them. It is in all these respects that an exemplar-based, a mechanistic and an interventionist approach, if properly construed, are thought to fit psychiatric disorders, and possibly to complement each other as approaches to scientific explanation. Other concerns, such as the need to decompose complex systems into subsystems or to keep a neutral attitude towards mind-brain matters can be more directly met by specific standpoints, either of an exemplar-based and mechanistic or interventionist kind respectively.

It is noteworthy that some years before separately defending the mechanistic and interventionist views, Kendler put forward a more – so to speak – "ecumenical" position. In a previous work of his, published in 2005, he explicitly suggested that explanatory pluralism is the most adequate approach to understand the nature of psychiatric illness, where explanatory pluralism is taken to acknowledge and interrelate perspectives focusing on distinct levels, thus recognizing the relevance of – in addition to neurobiological processes, genetic risk factors or neurochemical alterations – such elements as first person mental processes, cultural processes and psychological factors in the etiology of disorders. Epidemiological studies on the onset of major depression in twins, for examples, show that a severely stressful life, levels of loss, humiliation and entrapment, are predictors of depression. "Although humiliation is ultimately expressed in the brain, this does not mean that the basic neurobiological level is necessarily *the most efficient level* at which to observe humiliation" (Kendler 2005, p. 436), and, hence, to explain depression. The same sort of considerations work for cultural processes affecting other psychiatric illnesses, which will hardly be most effectively understood at the level of basic brain biology, such as the relation between bulimia and Western cultural models regarding body image. Explanatory pluralism is thus advocated to address *the most appropriate level in the given circumstances*.

Reflecting on the articulation of psychiatric explanation into levels, and questioning how they can affect the development of an etiologically based nosology, Kendler (2012b) has also pointed to seven criteria to evaluate how much weight should be given to an explanatory account: (i) *strength*: reflects the magnitude of the association between the explanatory variable and disease risk; (ii) *causal confidence*: reflects the degree to which the risk factors truly alter the probability of disease occurrence; (iii) *generalizability*: reflects the degree to which an explanation applies across a wide range of background conditions; (iv) *specificity*: refers to the degree to which the explanation applies only to the disorder under consideration; (v) *manipulability*: reflects the degree to which the identified risk factors can be modified by an intervention and how interventions can impact on the risk of the disease; (vi) *proximity*: refers to the location of the risk factor in the causal process underlying the disease; (vii) *generativity*: reflects the potential of the explanatory variables identified to gain further etiological understanding of the disease. Once again, the variety of diseases under examination must be taken into account, and not all these criteria apply equally well to all psychiatric disorders considered. While a pathology like, for instance, cystic fibrosis can be approached – to start with – via a gene-level analysis and can be well explained by mutations in the protein CF transmembrane conductance regulator gene (CFTR) according to *all* seven criteria, the situation is much more blurred in cases like alcohol dependence, which has no obvious candidate on which to base the nosology. Aldehyde dehydrogenase (ALDH) variants happen to be very highly evaluated in terms of strength, causal confidence, specificity and manipulability; social norms expectations and taxation, in turn, perform highly as explanations with respect to strength, causal confidence and specificity, very highly with respect to manipulability, and poorly or very poorly with respect to proximity and generativity. That not all criteria are met is by no

means to be regarded as a limit: the value attached to a given criterion for an adequate explanation depends on who is searching for an explanation and why. A basic behavioral researcher focusing on etiology will be looking for explanations with high levels of causal confidence, strength and proximity; a clinical researcher, interested in knowing how treatment of a certain condition would differ from treatment of others, will evaluate highly specificity and manipulability; people focusing on public health and prevention will deem important causal confidence, strength, generalizability and manipulability, and might even consider low specificity a virtue if aiming to reduce risks for a broad area of disorders. Therefore, different criteria of a good explanation will be advocated, according to the target levels and the purpose of enquiry, which are strictly connected.

Various aspects emerging from what has been presented above are worth highlighting. Pluralistic trends expressed in the literature referred to above tackle different issues, which tend to be conflated and partly confused:

(a) different sorts of explanations can be employed which identify causal factors at some specific level (e.g. neurobiological; psychological; socio-economic; ...); they are compatible and can be integrated with one another;
(b) different general conceptions of what "to explain" amounts to can be embraced in the search for psychiatric explanations (e.g. exemplar-based; mechanistic; interventionist; ...), which can be combined.

These two themes are related but distinct. Different kinds of variables, at different levels, play a causal role in the onset of psychiatric disorders, and different ways to conceive of and elaborate causal explanations can be adopted. Neurobiological, psychological, social, etc., accounts tend to privilege one level over the others; explanations focusing on variables at a single level can be elaborated according to different conceptions of scientific explanation, and the same account of explanation (e.g. mechanistic or interventionist) can consider many different explanatory levels at the same time. What seems to be one of the most pressing issues emerging from the examples illustrated in Sect. 1 and from the concerns expressed in some pluralistic proposals is that, while it is recognized that pathologies are multifactorial, explanations privileging one level over the others, or assuming or fuelling dichotomies, still persist in practice.

With respect to point (b), different approaches can and want to account for the multifactorial character of disorders, while dealing differently with other issues. In particular, while the interventionist approach is advocated in this context by insisting that it allows us to avoid ontological commitments – thereby simplifying the investigation to some extent – the search for mechanisms is typically associated with the specification of entities and activities, or the spatio-temporal processes involved. The merits of mechanistic approaches include their very capability to uncover specific kinds of entities and interactions. Productive relations to be represented by mechanistic systems in psychiatry, though, are often still quite opaque, and this typically leads to the identification of sketchy mechanisms rather than the elaboration of detailed mechanistic descriptions. For instance, we do not know how genetic factors interact with humiliation in the light of tragic childhood episodes in

order to bring about depression in an alcoholic, or whether and how this interaction is different from the interaction by which genes contribute to depression in a neurotic divorced man (see Murphy 2010, p. 603). A multiplicity of approaches can hence be considered, not to select *the* correct one, but to appreciate the distinctive contribution of each of them. The next section will further examine how explanatory pluralistic suggestions in psychiatry can be interpreted in the light of psychiatric concerns.

3 Psychiatric Explanations and Explanatory Pluralism

In presenting the kind of explanatory pluralism he believes should be adopted in psychiatry, Kendler appeals to the notion of *integrative pluralism*, seen as the approach that makes "active efforts [. . .] to incorporate divergent levels of analysis. This approach assumes that, for most problems, single-level analyses will lead to only partial answers. However, rather than building large theoretical structures, integrative pluralism establishes small 'local' integrations across levels of analysis". Psychiatry is considered to be in need of integrative pluralism insofar as scientists in the field "cross borders between different etiological frameworks or levels of explanation", working "bit by bit toward broader integrative paradigms" (Kendler 2005, p. 437). Pluralism can come into different versions, whose features and possible intersections deserve to be further specified.[9] If integrative pluralism is the guiding idea, it must be stressed that integrations cannot be achieved but by constructive interactions between different models, and by elaborating models accounting for the interactions among the different levels to which causal factors belong.

If we focus our attention on what we have seen so far, it appears that what specifically motivates explanatory pluralism here are the acknowledgment that causal variables belong to a multiplicity of levels and the purpose of avoiding reductionist positions; the aim of avoiding "the clumsy and out-dated baggage left from Cartesian dualism" (*id*., p. 439); and the conviction that an approach to explanation should be empirically-based. A guiding concern is that no level should be privileged a priori, as the most important, or the most scientific and hence most valuable. At the same time, embracing explanatory pluralism in psychiatry does not mean treating all methodologies as being of equal value. Can we better specify not only what we are pluralists about, but also what kind of pluralists we are in this context? Is there any underlying idea that some sort of complete explanatory picture can be – sooner or later – elaborated, or is some more radical form of pluralism advanced here? Is pluralism suggested here as only the acknowledgment of the existence and toleration of a diversity of current explanatory theories, or also as

[9]For reflections on various forms of pluralism and suggestions on tentative taxonomies, see Van Bouwel (2014).

the idea that distinctive views will persist as such in the long run?[10] In other terms, is actual plurality treated in this context as provisional and resolvable, or is the idea that renouncing pluralism would lead to some loss of explanatory information?

As emerges from the examples considered in Sect. 1, the same disorder can be dealt with by means of different models of explanation for different purposes; different sets of disorders can be examined from different perspectives; different models can be adopted for the same disorder by different professional roles. Pictures to be "filled in", "patchy reductions" leading to "piecemeal integration", and "bit-by-bit" efforts of integrative pluralism are suggested. "Such efforts should, over time, result in clarification of parts of the causal network from which it may be possible to move toward *a more complete etiological understanding* of the extremely complex mind-brain dysfunctions that it is our task to understand and treat" (*id.,* p. 436, italics added). On the one hand, it is maintained that "psychiatric disorders are, *by their nature*, complex multilevel phenomena" (*id*., p. 439), and some ideal "more complete understanding" is pursued. On the other hand, though, a deeper analysis of the "nature" of such pathologies is not the main concern of the works suggesting explanatory pluralism, which are under consideration here. Explanatory pluralism is not presented as a transient solution, and is mainly given an empirically-based and pragmatic justification.[11] It is acknowledged and stressed that, *as a matter of fact*, we are far from developing a full causal network for any psychiatric disorder, and that the wide range of different perspectives adopted deeply rely on contextual and pragmatic matters. What dictates that a given level or a few levels are to be picked out as the most significant from an explanatory point of view, and how they are identified, largely depend on the target we are aiming at.

Different focuses and targets pursued by different actors in psychiatric contexts are very unlikely to converge, nor should they do so: elements addressed by and factors motivating, e.g., basic behavioral researchers, clinical researchers and the public health service will not easily coincide, not even in the long run. A basic behavioral researcher will be looking for etiological explanations indicating, for instance, proximate causes; psychiatric epidemiologists focusing on public health and prevention will deem especially important causal factors that can be generalized and manipulated; a clinical psychiatrist, interested in treatments, will evaluate highly manipulable conditions as referred to the single case at stake. A single approach is unlikely to accommodate all explanatory interests and goals. Explanations can be employed to meet a range of different needs, such as enhancing theoretical understanding, predicting, preventing, controlling something's functioning by effectively intervening over it, designing experiments, setting a research agenda. "What we can best hope for is lots of small explanations, from a variety of explanatory perspectives, each addressing part of the complex etiological processes leading to disorders" (*id*., p. 435), and evaluated differently according to the extent to which they answer the various needs. The importance of the causal factors

[10]For some general reflections on plurality and pluralism, see Kellert et al. (2006).

[11]On pragmatic factors as a justification of pluralism see e.g. Waters (2006) and Giere (2006).

identified will be relative to the circumstances, the features of the specific pathology under consideration, the extent to which it has been understood so far, the resources available and the opportunities a given explanation provides.

Different etiological accounts can interact and complement each other in an ongoing process of investigation. Troubles seem to arise because, while in theory it is recognized that factors at many levels contribute to the onset of a disease and that they can be sought by a number of methodological tools, explanatory models employed in psychiatric practice often emphasize just one level, or a very limited set of levels, and fail to search deep into mutual interactions between the causal elements involved. As has been stressed, which causal factors to consider, and which to obscure, cannot but depend on the different epistemic interests an explanatory model is adopted for, and affect the investigative or practical utility of the model in different contexts. However, also when providing, for instance, only partial and sketchy mechanisms or when identifying a limited set of locally manipulable relations, etiological explanations can prove very useful. Etiologically partial models, describing psychiatric disorders with different – and, more often than not, low – degrees of resolution, can be adopted both for research and practice, e.g. to orient investigations, produce treatment benefits, guide health policies. "Psychiatric explanations are coherent and plausible insofar as they are *pragmatically useful* and *empirically testable* in clinical settings" (Brendel 2003, p. 569, italics added). While they can be mutually integrated to various extents, neither their unification nor their revealing some alleged ultimate nature of the disease are to be considered the primary goals of the discipline.

Nothing in pluralism dictates that all the perspectives are created equal. Some can be better than others in many respects, e.g. more adequate for some diseases than for others and more widely applicable by some professional roles. Pluralism is not in itself always good or always bad, and it is not *the* option to be embraced everywhere. Not even some specific integration of different approaches will do everywhere. Different models will continue to be used, and their use pragmatically justified, according to the object of investigation, context of inquiry and purposes of the investigator. A fruitful form of empirically-based, pragmatically justified and ontologically neutral pluralism will acknowledge that psychiatric disorders are such that "each factor is necessary for the phenomenon to have the various characters it has, but a complete account is not possible in the same representational idiom and is not forthcoming from any single investigative approach (as far as we know)" (Kellert et al. 2006, p. xiv). Instead of having just a plurality of unrelated views, though, a pluralistic attitude will be welcome and promoted insofar as researchers, clinicians, social carers and patients will become more and more *aware* that "for psychiatric disorders, explanatory power is dispersed and diffuse" (Kendler 2012b, p. 16). The promotion of an active integration between a multiplicity of perspectives does not amount to the neglect of some specific positions: if explanatory pluralism should be a desideratum in dealing with psychiatric disorders, specialists are likely consciously to continue to support a given model as the most adequate in the given circumstances. A geneticist or a pharmacologist, e.g., will not be required to turn into a psychiatric epidemiologist or a nurse – nor could they easily do

so – but to entertain various contributions as possibly complementing their own. This target can be pursued by direct dialogue and exchange between different professional roles, as well as by professional figures and patients; by clearly defining the questions, focuses and aims of explanation; and by making adopted models and underlying assumptions as explicit as possible, with no claim of superiority of one level of analysis over the others. Without renouncing field-specific accounts and stances, each in need of being defined on its own, pluralism shall be pursued to suggest fruitful integrations in the cases at stake. Pluralism shall be grounded on the recognition of the merits of different models, and shall not aim at eliminating diversity, but at gleaning the most from a plurality of available views.

4 Concluding Remarks

If, in general, "medicine is moving in the right direction but is not fully and truly explanatory pluralist yet" (De Vreese et al. 2010, p. 373), reflections on psychiatry – expressed from within the discipline and also with an eye on the philosophical scenario – have led to a few steps forward in that direction, and the idea that explanatory pluralism is the best option has been suggested as of paramount importance in the elaboration of a "philosophical structure for psychiatry" (see Kendler 2005).[12] In practice, a *plurality* of explanations are already adopted in psychiatry. That per se, though, does not make psychiatry explanatory *pluralist*. A broad acknowledgment of the different views available and an active confrontation between them is required.

What can *pluralism* be good for in the context of *psychiatric practice,* which seems to be increasingly expressing also theoretical concerns? To affirm the use of multiple approaches to psychiatric explanation and their complementing each other might help avoid the risk of a cacophony of views, with then difficult consensus on action and research trends; pluralism works as a reminder that there are many different ways in which mental illnesses can be dealt with and intervened upon, and promotes a genuine dialogue between different standpoints. The awareness that a huge number of factors endowed with explanatory power cross-talk to each other in psychiatric disorders and that they can be identified more or less easily and effectively by different approaches to explanation, and the recognition of the role of contextual matters should not be confined to theoretical inquiry on the topic, but increasingly promoted for effective prevention, treatment and decision-making in the field. Furthermore, it is worth recalling that a complete etiological picture might be useful, and even fruitfully striven for, but it is not necessary for good scientific practice. Being open to the possibility that, at least in principle, explanatory pluralism can be a *permanent* state allows to avoid relying on any single account, can lead to useful and not definitive partitioning of the field and decomposition

[12] See also Murphy (2010, p. 609) and Kendler (2012a, p. 385).

of specific disorders, and encourages case-by-case empirical analyses, without struggling for anything like *the true account* of psychiatric disorders. "As a practical discipline, psychiatry is concerned more with its methodology than its ontology: by adopting a pragmatic" – and, I shall add, genuinely pluralistic – "position on explanatory models, psychiatrists do not necessarily commit themselves to a particular view on the underlying structure of the universe" (Brendel 2003, p. 569).

Acknowledgments I am grateful to Jeroen Van Bouwel, Marcel Weber and Roberta Passione for their comments and suggestions on a draft version of this paper.

References

Andreasen, N. 1997. Linking mind and brain in the study of mental illnesses: A project for a scientific psychopathology. *Science* 275: 1586–1593.

Bechtel, W. 2007. *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Hillsdale: Lawrence Erlbaum Associates.

Black, K. 2005. Psychiatry and the medical model. In *Adult psychiatry*, 2nd ed, ed. E. Rubin and C. Zorumski, 3–15. Oxford: Blackwell.

Brendel, D.H. 2003. Reductionism, eclecticism, and pragmatism in psychiatry: The dialectic of clinical explanation. *Journal of Medicine and Philosophy* 28: 563–580.

Brog, M.A., and K.A. Guskin. 1998. Medical students' judgment of mind and brain in the etiology and treatment of psychiatric disorders. *Academic Psychiatry* 22: 229–235.

Campbell, J. 2008. Causation in psychiatry. In *Philosophical issues in psychiatry. Explanation, phenomenology, and nosology*, ed. K. Kendler and J. Parnas, 196–216. Baltimore: The Johns Hopkins Press.

Colombo, A., et al. 2003. Evaluating the influence of implicit models of mental disorder on processes of shared decision making within community-based multi-disciplinary teams. *Social Science & Medicine* 56: 1557–1570.

De Vreese, L., E. Weber, and J. Van Bouwel. 2010. Explanatory pluralism in the medical sciences: Theory and practice. *Theoretical Medicine and Bioethics* 31: 371–390.

Dick, D.M., and L.I. Bierut. 2006. The genetics of alcohol dependence. *Current Psychiatry Reports* 8: 151–157.

Edenberg, H.J., et al. 2004. Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations. *American Journal of Human Genetics* 74: 705–714.

Giere, R.N. 2006. Perspectival pluralism. In *Scientific pluralism*, ed. S.H. Kellert, H.E. Longino, and C.K. Waters, 26–41. Minneapolis: University of Minnesota Press.

Harland, R., et al. 2009. A study of psychiatrists' concepts of mental illness. *Psychological Medicine* 39: 967–976.

Kellert, S.H., H.E. Longino, and C.K. Waters. 2006. Introduction. The pluralist stance. In *Scientific pluralism*, ed. S.H. Kellert, H.E. Longino, and C.K. Waters, vii–xxix. Minneapolis: University of Minnesota Press.

Kendler, K. 2005. Toward a philosophical structure for psychiatry. *American Journal of Psychiatry* 162: 433–440.

Kendler, K. 2008. Explanatory models for psychiatric illness. *American Journal of Psychiatry* 165: 695–702.

Kendler, K. 2012a. The dappled nature of causes of psychiatric illness: Replacing the organic-functional/hardware-software dichotomy with empirically based pluralism. *Molecular Psychiatry* 17: 377–388.

Kendler, K. 2012b. Levels of explanation in psychiatry and substance use disorders: Implications for the development of an etiologically based nosology. *Molecular Psychiatry* 17: 11–21.

Kendler, K., and J. Campbell. 2009. Interventionist causal models in psychiatry: Repositioning the mind–body problem. *Psychological Medicine* 39: 881–887.

Kendler, K., and J. Parnas (eds.). 2012. *Philosophical issues in psychiatry II: Nosology*. Oxford: Oxford University Press.

Kendler, K., and C.A. Prescott. 2006. *Genes, environment, and psychopathology: Understanding the causes of psychiatric and substance use disorders*. New York: The Guilford Press.

Miresco, M.J., and L.J. Kirmayer. 2006. The persistence of mind-brain dualism in psychiatric reasoning about clinical scenarios. *American Journal of Psychiatry* 163: 913–918.

Molina, J.C., et al. 2007. The International Society for Developmental Psychobiology 39th annual meeting symposium: Alcohol and development: Beyond fetal alcohol syndrome. *Developmental Psychobiology* 49: 227–242.

Murphy, D. 2010. Explanation in psychiatry. *Philosophy Compass* 5(7): 602–610.

Murphy, D. 2011. Conceptual foundations of biological psychiatry. In *Philosophy of medicine*, ed. F. Gifford, 425–451. London: Elsevier.

O'Connor, R.M., T.G. Dinan, and J.F. Cryan. 2012. Little things on which happiness depends: MicroRNAs as novel therapeutic targets for the treatment of anxiety and depression. *Molecular Psychiatry* 17: 359–376.

Patil, T., and J. Giordano. 2010. On the ontological assumption of the medical model of psychiatry: Philosophical considerations and pragmatic tasks. *Philosophy, Ethics, and Humanities in Medicine* 5: 3.

Schaffner, K. 2002. Clinical and etiological psychiatric diagnoses: Do causes count? In *Descriptions and prescriptions: Values, mental disorders and the DSMs*, ed. J. Sadler, 271–290. Baltimore: The Johns Hopkins University Press.

Van Bouwel, J. 2014. Pluralists about pluralism? Different versions of explanatory pluralism in psychiatry. In *New directions in the philosophy of science*, ed. M.C. Galavotti, D. Dieks, W.J. Gonzalez, S. Hartmann, Th. Uebel, and M. Weber. Dordrecht: Springer.

Waters, C.K. 2006. A pluralist interpretation of gene-centered biology. In *Scientific pluralism*, ed. S.H. Kellert, H.E. Longino, and C.K. Waters, 190–214. Minneapolis: University of Minnesota Press.

Woodward, J. 2008. Cause and explanation in psychiatry. In *Philosophical issues in psychiatry. Explanation, phenomenology, and nosology*, ed. K.S. Kendler and J. Parnas, 132–184. Baltimore: The Johns Hopkins Press.

Pluralists About Pluralism? Different Versions of Explanatory Pluralism in Psychiatry

Jeroen Van Bouwel

1 Introduction

In her paper *Explanatory Pluralism in Psychiatry: What Are We Pluralists About, and Why?* Raffaella Campaner presents a strong defense of explanatory pluralism in psychiatry with a primary or main emphasis on explanatory pluralism as opposed to explanatory reductionism, be it biological, psychological or social reductionism. She thus eschews hard reduction as well as hard emergence (cf. Kendler 2008, p. 700). In this contribution, I briefly revisit some of Campaner's examples of plurality and pluralism in psychiatry (Sect. 2) and then I shift the focus to the variety of understandings of explanatory pluralism, explicating different versions of pluralism (Sect. 3). In Sect. 4, I discuss the pros and cons of these different versions of explanatory pluralism. Finally, in Sect. 5, I raise the question of how to implement or operationalize explanatory pluralism in scientific practice; how to structure the "genuine dialogue" or shape "the pluralistic attitude" that Campaner is referring to in her paper. The overall aim of my contribution is to shift the focus from pluralism as a critique of reductionism towards analyzing the different existing versions of pluralism in science and how to implement them.

J. Van Bouwel (✉)
Centre for Logic and Philosophy of Science, Ghent University,
Blandijnberg 2, 9000 Ghent, Belgium
e-mail: Jeroen.VanBouwel@UGent.be

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_8,

2 Plurality in Psychiatric Practice and the Challenges It Poses

Analyzing plurality in psychiatry, Campaner starts with discussing several studies that have been performed about how models of psychiatric disorders are actually employed in medical practice by trainee psychiatrists, medical students, health professionals, service users and practitioners in psychiatry (cf. Harland et al. 2009; Brog and Guskin 1998; Colombo et al. 2003, and, Miresco and Kirmayer 2006). The different approaches to studying mental disorders are labeled as biological, cognitive, behavioral, psychodynamic, social, etc. These approaches are playing on *different levels,* within *different disciplinary fields,* e.g., epidemiology, psychology, neurology, genetics, or socio-economic inquiries, involving *different kinds of evidence*, e.g., biomolecular research, epidemiological studies, or first-person reports, and, *different focuses*, e.g., on populations, subclasses of populations, or individuals.

The studies teach us something about the existing plurality of models in psychiatry and the variety in use among current and future practitioners. We learn that even though the biological model might be the most strongly endorsed, model endorsement varies with disorder considered, so there is no exclusive commitment to any one model. Further, different health professionals and service users, i.e. psychiatrists, community psychiatric nurses, social workers, patients and informal caregivers, embrace different etiological models of disorders. In general, psychiatrists and psychiatric nurses were more in favor of the medical approach, while social workers in general tended to endorse the social model. Among patients there was a higher heterogeneity. The studies also show how the old mind-brain dichotomy is still alive and well, often unacknowledged, as well as the impact it has on attributions of personal responsibility.

The challenges this plurality poses for psychiatry are manifold. Adopting one of the models of disorders without being aware of it does seem to be far from optimal and might have dire implications. Employing different implicit explanatory models might, for instance, lead to conflicting assumptions that create misunderstandings (in communication, diagnosis and treatments) among psychiatric practitioners and other professionals in health care, as studies show (e.g., Colombo et al. 2003).

A first important step is then to make the different implicit models explicit. Given the real implications they have, undoubtedly informing diagnosis, treatments, prevention strategies and other substantial decisions, this is crucial. Making the implicit models explicit will increase awareness of the different models at play, improve communication and lead to more consistency in dealing with mental health problems in practice.

Secondly, once the different models at play are made more explicit and users are aware of the existing plurality, the challenge is how to deal with the plurality in the best possible way – a challenge for both researchers and practitioners. Does the

plurality have to be – and can it be – resolved? If not, and we would normatively endorse plurality and advocate pluralism, then how should that pluralism be understood or characterized? And how could it be implemented? Those are the questions that will be addressed in the following sections.

3 Different Ways of Dealing with Plurality – Contending Versions of Pluralism

3.1 Explanatory Pluralism Versus Explanatory Reductionism

Campaner addresses the questions just raised by discussing several pluralistic stances put forward in the elaboration of models of diseases and their explanation. She focuses in particular on the work of Dominic Murphy, Kenneth Kendler and John Campbell. Their accounts of explanation have several aspects in common, as Campaner notes, namely that they include multiple explanatory levels, avoid exclusive reduction to the biological level and acknowledge the irreducible role higher level explanations might play. This is one aspect of Campaner's characterization of explanatory pluralism, namely: "(a) different sorts of explanations can be employed which identify causal factors at some specific level (e.g. neurobiological; psychological; socio-economic; . . .); they are compatible and can be integrated with one another." (Campaner 2014) The second aspect of explanatory pluralism that Campaner highlights, is: "(b) different general conceptions of what "to explain" amounts to can be embraced in the search for psychiatric explanations (e.g. exemplar-based; mechanistic; interventionist; . . .), which can be combined." Thus, Kendler's mechanistic account of explanation, Murphy's exemplar-based account and Kendler and Campbell's interventionist account can all be embraced and combined. The pluralist can emphasize that each of these accounts of explanation has its specific strengths and capabilities, as Campaner illustrates.

Thus, this characterization of explanatory pluralism shows how it is clearly opposed to explanatory reduction. First, there is no a priori privileged level of explanation. For instance, even though mental disorders might ultimately be expressed in the brain, the neurobiological level is not necessarily always the most appropriate level at which to explain a disorder. Second, there is not one correct way of providing explanation that should be the standard for all explanations; different general conceptions of explanations, e.g. mechanistic, interventionist, exemplar-based, should be considered and appreciated for their respective strengths. Campaner articulates what the advocates of explanatory pluralism have in common and contrasts it with reductive approaches. Next, she raises a question that might be the starting point to discuss different versions of pluralism, starting to highlight differences among explanatory pluralists, between understandings of explanatory pluralism.

3.2 *Different Understandings of Explanatory Pluralism*

Campaner raises the question whether the existing plurality is: (1) considered as eventually resolvable, i.e. explanatory pluralism is "only the acknowledgement of the existence and toleration of a diversity of current explanatory theories" and in the long run a complete explanatory picture will emerge, or, (2) is it rather "the idea that distinctive views will persist in the long run" and that a single, complete explanatory picture is very unlikely to emerge? In this Section, I elaborate that it would be helpful to distinguish more than those two versions of pluralism, i.e. more than two different understandings of explanatory pluralism both within philosophy of science and psychiatry. I distinguish five different versions, namely *moderate/temporary pluralism, anything goes pluralism, isolationist pluralism, integrative pluralism* and *interactive pluralism.* The first four of them are discussed by Sandra Mitchell (2009) and the last version of pluralism is mine.[1]

Moderate pluralism advocates to "recognize and promote a temporary plurality of competing theories as means toward achieving unity of science in the long run." (Mitchell 2009, p. 108). It is this version of plurality and pluralism Campaner refers to as temporary and resolvable. Mitchell herself cannot subscribe to this form of pluralism, as it eventually wants a single, true unified theory (a monist goal), and this does not dovetail with the ontology of complex systems in which the multilevel structure encourages focused analysis at each level.

Next, Mitchell distinguishes *anything goes pluralism* that represents "the advocacy of retaining all, possibly inconsistent, theories that emerge from a community of investigators." (Mitchel 2004, p. 85) Just like reducing a collection of analyses of the same phenomenon to one single model or theoretical framework, Mitchell finds retaining all theories that emerge equally unacceptable and not supported by actual scientific practice (Mitchell 2009, p. 108). Instead, she wants to explore the middle ground between monism and anything goes where she distinguishes integrative pluralism from isolationist pluralism.

Mitchell herself advocates *integrative pluralism* (cf. Mitchell 2002, 2004, 2009).[2] Integrative pluralism takes into account both today's highly specialized (sub)disciplinary research and the need of integrating the respective findings concerning a phenomenon: "Developing models of single causal components, such as the effects of genetic variation, or of single-level interactions, such as the operation of selection on individuals (. . .) need to be integrated in order to understand what historical, proximal, and interactive processes generate the array of biological phenomena we observe. Both the ontology and the representation of complex systems

[1] For other taxonomies of pluralism, see, e.g., Kellert et al. (2006) and Van Bouwel (2009).

[2] It should be noted that the idea of (the possibility of) integration also appears in the first part (a.) of Campaner's characterization of explanatory pluralism (see Sect. 3.1, above). Second, one of the authors discussed by Campaner, Kenneth Kendler, approvingly refers to Mitchell's integrative pluralism (cf. his 2005, p. 437).

recommend adopting a stance of integrative pluralism, not only in biology, but in general." (Mitchell 2004, p. 81). However complex, and however many contributing causes participated, there is only one causal history that, in fact, has generated a phenomenon to be explained. Thus, according to Mitchell's integrative pluralism, "it is only by integration of the multiple levels and multiple causes (. . .) that satisfactory explanations can be generated." (Mitchell and Dietrich 2006, p. S78)

Mitchell opposes her integrative pluralism to *isolationist pluralism* or *"levels of analysis" pluralism*. According to this understanding of explanatory pluralism different questions invoke different explanatory schemata, and there is no need to consider explanations developed at levels other than their own or for intertheory relations among the levels. This limits the interaction between various theories offering explanations in a given domain and leads to isolation, according to Mitchell. "If there is no competition between levels, there need be no interaction among scientists working at different levels either. The problem with the isolationist picture of compatible pluralism is that it presupposes explanatory closure within each 'level of analysis' and a narrowness in scope of scientific investigation that precludes the type of fruitful interactions between disciplines and subdisciplines that has characterized much of the history of science." (Mitchell 2004, p. 85)

There is (at least) one possible understanding of pluralism that Mitchell does not discuss and that I want to introduce here. Let us label it *interactive pluralism*. It is situated in between integrative and isolationist pluralism, as: (a) on the one hand, it claims that satisfactory explanations can also be obtained without integrating of multiple levels, so there is no integration imperative, and, (b) on the other hand, it does not discourage interaction as, in some instances, interaction and integration do lead to better explanations.

Placed on a continuum going from monism to *anything goes* pluralism, we thus have monism, moderate pluralism, integrative pluralism, interactive pluralism, isolationist pluralism and *anything goes* pluralism. This ordering reflects increasing strength of the pluralist position. All five versions of what explanatory pluralism is or should be will answer differently on the questions raised at the end of Sect. 2 – and therefore it is important to go beyond the two versions of pluralism articulated by Campaner. In the next Section, I will raise some questions about the three versions of pluralism that cover the middle ground between moderate/temporary pluralism and *anything goes* pluralism. This also gives us the opportunity to articulate some important differences.

4 Questioning and Evaluating the Different Understandings of Explanatory Pluralism

Having spelled out different possible understandings of explanatory pluralism, I would now like to discuss the question of whether any of these versions of pluralism is more convincing than the other ones. Below, I briefly raise some

challenges concerning integrative and isolationist pluralism, and emphasize the benefits of interactive pluralism.[3]

4.1 Questioning Integrative Pluralism

A first question concerning integrative pluralism asks whether integration is always necessary to obtain a "satisfactory explanation", as Mitchell claims. Straightforward reduction might sometimes lead to very satisfactory explanations efficiently serving our explanatory interest (cf. Van Bouwel et al. 2011).[4] Integration might very well be a good heuristic advice or play a justificatory role, but why should it be a criterion for a satisfactory explanation?

Second, won't integrated explanations often provide us with too much information and therefore be less efficient in providing the answers we are looking for, in answering our explanation-seeking questions?[5] In his book *The Rise and Fall of the Biopsychosocial Model*, Nassir Ghaemi (2010), discusses how this model for psychiatry included the idea that adding and integrating "more perspectives is always better". Eventually the approach was made unfeasible and uninteresting in practice by being too general and too vague. A similar evaluation has, for instance, been made about the developmental systems approach in studying human behavior (cf. Longino 2013). Integrative pluralism insufficiently acknowledges that explanations are always a trade-off between generality and preciseness, simplicity and realism, accuracy and adequacy, etc., depending on one's explanatory interests. Integrative explanations might be sometimes far too cumbersome, less efficient, and less adequate than possible alternative explanations.

Third, could the demand for integrated explanations not lead to losing *idioms/adequacy* in light of our explanatory interests, thus losing the capacity of answering some explanation-seeking questions in the most adequate way (*i.a.* strengthening hermeneutical injustice)?

Fourth, what would the integration imperative imply for *heterodox,* non-mainstream theories? What is the impact on the dynamics between research approaches? Think in particular about situations in which there is epistemic

[3]For a more extensive discussion and evaluation of different versions of pluralism, also see Van Bouwel (2009) and Van Bouwel (2014).

[4]I use (a trade-off between) *accuracy*, *adequacy* and *efficiency* here as criteria to evaluate what is a *satisfactory explanation*; (a) *accuracy* concerns the relation with reality, precise description, (b) *adequacy* refers to what the explainee expects from the explanation addressing the explanatory interest, and (c) *efficiency* points at the amount of work and/or information needed for the explanation (also see Sect. 5.1, below).

[5]Note that the use of *efficiency* as a criterium is also present in Kendler's work: "Although humiliation is ultimately expressed in the brain, this does not mean that the basic neurobiological level is necessarily *the most efficient level* at which to observe humiliation" (Kendler 2005, p. 436).

inequality, in which one research program at one level is a lot bigger and more elaborated than another one at another level and where integration risks minimizing dissent, overlooking diversity, eliminating differences and/or a homogenization in terms of the bigger one.[6]

4.2 Questioning Isolationist Pluralism

A first question that should be raised concerning isolationist pluralism, is: Does isolation always lead to better explanations? And, second, how can we know given the lack of competition between explanations coming from different approaches within this version of pluralism? According to Mitchell's characterisation of this position, the idea that some questions are better answered on one level and others on another leads to an isolationist stance with respect to the separate questions. Now, if there is no interaction or no intention of competition between levels, then there need be no interaction between scientists working at different levels either. Thus, this form of pluralism does not do much more than acknowledging plurality; it does not suggest any way of making the plurality epistemically as productive as possible.

Third, why do isolationist pluralists presuppose that interaction cannot be productive, while it is evident that fruitful interactions between (sub)disciplines have characterized much of the history of science as Mitchell mentions?

Fourth, as concerns the dynamics between research approaches, isolation, a lack of interaction between the mainstream/orthodoxy and the heterodoxy, e.g. in economics, seems to create a very static, non-productive situation in which, on the one hand, the traditional heterodoxy is aiming to become the new monist, the new mainstream, substituting the current orthodox one, while on the other hand, the orthodoxy or mainstream considers the heterodoxy as a constitutive outsider that proves the scientific status of the orthodoxy or mainstream (cf. Van Bouwel 2009).

[6]I think it is important to pay attention to the dynamics between different approaches in scientific practice. Mitchell does not pay enough attention to this aspect in defending her integrative pluralism. As I argued before (cf. Van Bouwel 2013, p. 417), given that reductionism is one of the main targets of Mitchell (2009)'s work, it might be insightful to study *all* possible factors at play in sustaining reductionist research (e.g., genetic research in the health business) rather than nonreductionist alternatives, like environmental health research; it might not merely be because of the wide-spread spirit of Newtonianism that reductionism still flourishes! Moreover, if Mitchell wants to plead for more nonreductionist research in combination with the integration imperative (very likely benefitting the bigger players), it seems indispensable to understand the role of values in the selection and formulation of research questions as well as how to foster valuable alternatives to the mainstream research programs.

4.3 Questioning Interactive Pluralism

Interactive pluralism, the possibility not discussed by Mitchell, might be a third option that avoids some of the worries about integrative and isolationist pluralism. Why?

First, where there is a presumption of reconcilability in integrative pluralism, and irreconcilability in isolationist pluralism, interactive pluralism considers the ir-/reconcilability to be an open question. In-depth analyses of scientific practice teach us that competing approaches often do not parse causal space in the same way (cf. Longino 2013). This is problematic for Mitchell's advocacy of integrative pluralism and its presumption of reconcilability in integrating multiple approaches in order to obtain the (one) causal history of the phenomenon to be explained.

Second, interactive pluralism questions whether integration would always lead to a better explanation as well as whether integration is necessary to obtain a "satisfactory explanation". As concerns the former, integrative explanations might sometimes be too general, vague and cumbersome, i.e., not always the most efficient. Mitchell does not take into account the adequacy and efficiency criteria in stipulating what is the most satisfactory explanation. As concerns the latter claim that integration would be necessary to obtain a satisfactory explanation, I mentioned above that we should rather consider the trade-off between accuracy, adequacy and efficiency of explanations in labelling what is "satisfactory". Always focusing on integration, irrespective of one's precise explanatory aims and needs in a given context, would – if even possible – unnecessarily complicate matters and even paralyze research and decision-making.

Third, even though integration is not imperative, interactive pluralism rejects isolation and endorses interaction and engagement, be it without the presumption of always reaching a consensus or an integration. Some (but definitely not all!) explanation-seeking questions might require a combination, integration or cooperation of models in order to address our explanatory interests as well as possible. The respective explanation-seeking questions can be channels of interaction between competing research programs. The interaction does not have to lead to integration, it might just help to refine the respective approaches as well as articulate the strengths and limitations of each of them.

Fourth, contrary to integrative pluralism, the mainstream and non-mainstream approaches start on equal footing. Even for heterodox approaches that cannot be easily integrated, the interaction with orthodox or other heterodox approaches is endorsed, because approaches are sharpened as a response to challenge and criticism, methodologies refined, concepts clarified, etc. Moreover, the interaction between explanatory approaches might also make the limitations of each approach evident by the articulation of questions that they are not designed to answer.

5 Philosophical Frameworks for Explanatory Pluralism

In Sect. 2, I mentioned some of the forms of plurality one encounters in psychiatry as well as the problems that may cause. The challenge is to find productive ways to deal with this plurality. Campaner talks of promoting "a genuine dialogue between different standpoints" as well as "a pluralistic attitude". It raises the question of how to implement or operationalize explanatory pluralism in scientific practice; how to structure a "genuine dialogue" or shape "the pluralistic attitude"? A discipline might show plurality while all the individual researchers (or practitioners) are monist. Is the discipline in that case really subscribing to explanatory pluralism, making the best of the existing plurality? I do not think so. Therefore, in this last section, I would like to offer some philosophical tools or frameworks that might be helpful in implementing pluralism. First, on the basis of my research, mainly concerning explanatory pluralism in the social sciences, I have developed a framework for understanding explanatory pluralism which can help to elaborate some of the points made by Campaner about explanatory pluralism, as I will argue in Sect. 5.1. Second, I suggest that another way to get more concrete about what a "genuine dialogue" would look like are Helen Longino's CCE-norms for critical interaction, which I will discuss in Sect. 5.2.

5.1 A Framework for Explanatory Pluralism

On the basis of my analysis of actual scientific practice, mainly in the social sciences, I developed a framework for understanding explanatory plurality in scientific practice (see, e.g., Van Bouwel and Weber 2002; Weber and Van Bouwel 2002). The framework works as a tool to (a) *make the explananda as explicit as possible*, and (b) pay attention to the *underlying explanatory, epistemic interests*. This is imperative for clarifying discussions about competing explanations: there are many cases where two explanations of the same phenomenon are perceived as competitors, but actually have different *explananda*. The framework employs the erotetic model of explanation that regards explanations as answers to why-questions. Making the explananda as explicit as possible as well as paying attention to the different epistemic interests, can be achieved by explicating the explanation-seeking questions and their logic.

Analyses of explanatory practice in science teach us that different explanation-seeking questions or requests should be distinguished. I do not consider the questions and motivations mentioned here as the only possible ones, but I do believe they are omnipresent in scientific practice. At least five types of explanatory questions can be distinguished:

(E) Why does *x* have property *P,* rather than the expected property *P'?*
(I) Why does *x* have property *P,* rather than the ideal property *P'?*

(I') Why does *x* have property *P,* while *y* has the ideal property *P'?*
(F) Is the fact that *x* has property *P* the predictable consequence of some other events?
(H) Is the fact that *x* has property *P* caused by a familiar pattern or causal mechanism?

First, explanation-seeking questions can require the explanation of a contrast, e.g., of the form (E), (I) and (I'). Contrastive (E)-type questions, for instance, can be motivated by surprise: things are otherwise than we expected them to be and we want to know where our reasoning process failed (which causal factors did we overlook?). Contrastive questions of type (I) and (I') can be motivated by a therapeutic or preventive need; they request that we isolate causes which help us to reach an ideal state that is not realised now, comparing the actual fact with the one we would like to be the case (therapeutic need) or to prevent the occurrence of similar events in the future (preventive need).

The form of a *contrastive* explanation (i.e., an answer to a contrastive question) enables us to obtain information about the features that differentiate the actual causal history from its (un)actualized alternative, by isolating the causes that make the difference. This information does not include information that would also have applied to the causal histories of alternative facts.

Second, non-contrastive explanation-seeking questions, concerning plain facts, like (F) and (H), are also omnipresent in science. These non-contrastive questions can have different motivations. One possible motivation is sheer intellectual curiosity, with a desire to know how the fact "fits into the causal structure of the world" or to know how the fact was produced from given antecedents via spatio-temporally continuous processes. A more pragmatic motivation is the desire for information that enables us to predict whether and in which circumstances similar events will occur in the future (or the anticipation of actions of persons/groups). Another possible motivation concerns causally connecting object *x* having property *P* to events we are more familiar with.

The form these explanations of *plain facts* (answers to non-contrastive questions) have, shows how the observed fact was actually caused, which implies providing the detailed mediating mechanisms in a (non-interrupted) causal chain across time, ending with the explanandum. Alternatively, answering to the second motivation, the explanation can follow a covering law/law-based model.

By making the different possible explanation-seeking questions explicit, the motivation – explanatory interest – and the explanatory information required will be taken into account. Given that one phenomenon can be the subject of different questions, and that we want to answer these different kinds of explanatory questions in *the best possible way,* different forms of explanation are indispensable. In order to decide on *the best possible way*, we consider (trade-offs between) the criteria (a) *accuracy* or relation with reality, precise description, (b) *adequacy* in relation to what the explainee expects from the explanation addressing the explanatory interest, and (c) *efficiency* or amount of work and/or information needed for the explanation. To clarify these criteria and the idea that there often is a trade-off

between them, let us compare explanations with maps. A subway map like the one of the Paris Metro is *adequate* for its users because it *accurately* represents specific types of features (e.g. direct train connections between stations, number of stations between two given stations, ...) while other features are deliberately represented *less accurately* (the exact distances between the stations, the relative geographical orientation of the stations, ...). If the latter would be represented more accurately, the map could become less *adequate* for its intended users and a perfectly accurate representation mirroring every detail would be utterly useless. Furthermore, one could make the map more accurate, less adequate (without being completely inadequate), but also a lot less *efficient* in use (e.g. by making it less abstract, providing more cumbersome, obsolete information or by being too demanding or complicated to use). Other maps (e.g. Paris' shopping or tourist attractions maps) require other kinds of information (relating to, e.g., distances, details about street names, house numbers, etc.) in order to be useful – the best trade-off between accuracy, adequacy and efficiency differs depending on the interests or desiderata at play. Thus, on the one hand, because of different interests or desiderata, it is impossible to make a map that is ideal in all possible situations. On the other hand, not all maps are equally good, as one can make claims of superiority that are bound to specific situations. The same can be said for forms of explanation.[7]

To sum up, an explanation is an answer that should be evaluated in relation to a question that is a specific request for information. The precise meaning of the question is therefore important. Making the explanation-seeking questions as explicit as possible may show that, given that explanatory interests and contexts select distinct objects of explanation, a phenomenon can be subject of very different explanation-seeking questions. Consequently, different answers/explanations are required in which the most *accurate*, *adequate* and *efficient* explanatory information (in relation to the explanatory interest) is provided. Thus, different forms of explanation on different levels are indispensable to answer the respective explanation-seeking questions in the best possible way.

Returning to the plurality discussed by Campaner, a framework such as the one just presented explicating the logic of explanation-seeking questions is a way to compare competing explanations and to raise awareness about plurality. Different models are helpful in addressing different questions; one model may describe some facets extremely well, while making abstraction of, or even distorting, other facets – facets that might be the focus of other models. Explaining why person P is alcohol-dependent, for instance, might then lead to distinguishing explanations-seeking questions such as: (a) Why is person P addicted to alcohol, while person Q, who also drinks alcohol regularly, is not?; (b) Why does person P drink 10 units of alcohol per day, rather than 2 units?; (c) Was person P's alcohol addiction predictable?; (d) Are people like person P often addicted to alcohol?[8] Besides making the differences

[7] Also see Van Bouwel and Weber (2008) for more about these criteria.

[8] More examples related to medical sciences and using the framework for explanatory pluralism can be found in De Vreese et al. (2010).

between explication-seeking questions explicit and, as such, helping to see what different kinds of causal information are required, the framework also highlights the different epistemic, explanatory interests underlying the explanation-seeking questions, be it prevention and public health, individual therapy, curiosity, etc. Different actors in the psychiatric context, e.g., clinical researchers, basic behavioral researchers, public health services, etc, have different interests and motivations, looking for different information that can be found in the most accurate, adequate and efficient way possible in different models. It is very unlikely that one and the same model would always be the most accurate, adequate and efficient given all (current and future) epistemic, explanatory interests. This makes plurality an epistemic virtue.

Finally, using this framework for explanatory pluralism does enable the dialogue and addresses the need for integration prominent in the literature on explanatory pluralism in psychiatry. However, integration is not understood here as a requirement on the level of the explanation, as an imperative to integrate explanations, but rather on a meta-level as agreeing about how to disagree or how to spell out disagreement within a common framework. Making the explanation-seeking why-questions and their underlying epistemic interests explicit, this framework helps to stipulate the strengths and weaknesses of the respective conceptions of explanation and levels of explanation in answering the explanation-seeking why-questions while taking into account (the trade-offs between) the criteria of accuracy, adequacy and efficiency.[9]

5.2 Framing the "Genuine Dialogue"?

A second way in which we can explicate the idea of a "genuine dialogue" and further the implementation of pluralism in science, consists in stipulating norms that guide the interaction among competing approaches or competing models of mental disorder in psychiatry**.** The finality of these norms is not so much to arrive at one integrated model of mental disorder, but rather to enable interaction that might sometimes lead to local integration, but might also lead to a clearer articulation of differences among models. Thus, there is no imperative for integration, but rather an imperative to interact and learn from each other, without losing the strengths of one's own angle or approach. Certain norms can frame the interaction as a meta-consensus or meta-agreement within which disagreement and plurality can flourish.

Helen Longino's (2002, pp. 128–135) four norms, for instance, might be considered as framing a dialogue among competing scientific approaches, organizing

[9]For more on our approach to scientific explanation, see Weber et al. (2013). Let me also mention that this approach fits well with Interactive Pluralism (however, developing this point as well as the relation of the framework to the other versions of pluralism, goes beyond the scope of this paper).

a framework for critical interaction. Although these norms are rather vague, they might be a good starting point:

1. *Venues for criticism.* There must be publicly recognized forums for the criticism of evidence, of methods, and of assumptions and reasoning. This norm also warns for the limitations of forums, e.g. because of commercial interests.
2. *Uptake of criticism.* Response and change, i.e. "the community must not merely tolerate dissent, but its beliefs and theories must change over time in response to the critical discourse taking place within it." (*id*., 129).
3. *Public standards.* This norm ensures that critical discourse is nonarbitrary; the standards regulate discursive interaction, and as they are public, not just implicit, they help both defenders of a certain claim and their critics to identify their points of agreement and disagreement and structure the process in which problems are handled. Longino adds that these standards are not static, but may themselves be criticized and transformed.
4. *Tempered equality of intellectual authority.* The community must be characterized by equality of intellectual authority, a norm that warns that social, political, and economic privilege and power ought not determine epistemic privilege and power. This norm is meant to impose duties of inclusion.

Adding these norms to the set of methodological norms in science enables a productive dialogue among the plurality of approaches and is conducive to:

- Criticizing background assumptions from a variety of perspectives, making the assumptions of an approach visible; values and interests are not eliminated or purified, but are addressed by more and different values and interests;
- Sharpening the investigative resources proper to each approach as a response to challenge and criticism, refinement of methodologies, clarification of concepts, . . . ;
- Explicating the limitations of each approach by the articulation of questions that they are not designed to answer; the limited range of an approach's concepts and methods, by making their respective epistemic interests or values explicit, etc.; rival approaches – depending on different concepts, methods, etc. – are shown to have empirical successes as well, be it in relation to other questions or driven by other interests and values;
- Providing a forum for capable contenders of the orthodoxies, the mainstream approaches.

Moreover, the dialogue has no a priori commitment to monism or integration; maintaining the possibility of alternative rules of data collection (including standards of relevance and precision), inference principles, epistemic interests, values and aims of inquiry.

Longino's account is one way in which the "genuine dialogue" might be framed and plurality might be made as productive as possible. I hope future philosophical research will focus more on this kind of approaches to plurality – be it to refine Longino's account or to develop fruitful alternatives that will help implement pluralism.

6 Conclusion

In this contribution, I, first, wanted to distinguish different versions of explanatory pluralism that exist in the literature, as well as discuss some of the pros and cons of these different versions. Explicating five different versions of pluralism elaborates on Campaner's distinction of two different versions of pluralism, one being explanatory pluralist and the other, implicitly, explanatory reductionist.

Second, in my evaluation of the different versions of pluralism, I raised several critical questions concerning *Integrative Pluralism* – a version of pluralism that made its way into the literature on explanatory pluralism in psychiatry (cf. Kendler 2005). *Interactive Pluralism* was presented as an alternative understanding of pluralism that does not have the problematic features *Integrative Pluralism* or *Isolationist Pluralism* have.

Third, I pointed at some of the problems that plurality engenders in practice, both in research and clinical practice, but also at the epistemic virtues of plurality. In order to make plurality as virtuous as possible in practice and advance pluralism, we should develop philosophical tools that help us with the implementation of pluralism. I suggested that my question-based framework for explanatory pluralism as well as Helen Longino's social-epistemological procedures for interaction might be interesting points of departure and show us a fertile future direction for philosophy of science in its dealing with plurality.

References

Brog, M.A., and K.A. Guskin. 1998. Medical students' judgment of mind and brain in the etiology and treatment of psychiatric disorders. *Academic Psychiatry* 22: 229–235.

Campaner, R. 2014. Explanatory pluralism in psychiatry: What are we pluralists about, and why? In *New directions in the philosophy of science*, The philosophy of science in a European perspective series, ed. M.C. Galavotti, D. Dieks, W.J. Gonzalez, S. Hartmann, T. Uebel, and M. Weber. Berlin: Springer.

Colombo, A., et al. 2003. Evaluating the influence of implicit models of mental disorder on processes of shared decision making within community-based multi-disciplinary teams. *Social Science & Medicine* 56: 1557–1570.

De Vreese, L., E. Weber, and J. Van Bouwel. 2010. Explanatory pluralism in the medical sciences: Theory and practice. *Theoretical Medicine and Bioethics* 31: 371–390.

Ghaemi, N. 2010. *The rise and fall of the biosychosocial model*. Baltimore: The Johns Hopkins University Press.

Harland, R., et al. 2009. A study of psychiatrists' concepts of mental illness. *Psychological Medicine* 39: 967–976.

Kellert, S.H., H.E. Longino, and C.K. Waters. 2006. Introduction. The pluralist stance. In *Scientific pluralism*, ed. S.H. Kellert, H.E. Longino, and C.K. Waters, vii–xxix. Minneapolis: University of Minnesota Press.

Kendler, K. 2005. Toward a philosophical structure for psychiatry. *American Journal of Psychiatry* 162: 433–440.

Kendler, K. 2008. Explanatory models for psychiatric illness. *American Journal of Psychiatry* 165: 695–702.

Longino, H.E. 2002. *The fate of knowledge*. Princeton: Princeton University Press.
Longino, H.E. 2013. *Studying human behavior*. Chicago: University of Chicago Press.
Miresco, M.J., and L.J. Kirmayer. 2006. The persistence of mind-brain dualism in psychiatric reasoning about clinical scenarios. *American Journal of Psychiatry* 163: 913–918.
Mitchell, S. 2002. Integrative pluralism. *Biology and Philosophy* 17(1): 55–70.
Mitchell, S. 2004. Why integrative pluralism? *E:CO* 6(1–2): 81–91.
Mitchell, S. 2009. *Unsimple truths. Science, complexity, and policy*. Chicago: University of Chicago Press.
Mitchell, S., and M. Dietrich. 2006. Integration without unification: An argument for pluralism in the biological sciences. *The American Naturalist* 168: S73–S79.
Van Bouwel, J. 2009. The problem with(out) consensus: The scientific consensus, deliberative democracy and agonistic pluralism. In *The social sciences and democracy*, ed. J. Van Bouwel, 121–142. Basingstoke: Palgrave Macmillan.
Van Bouwel, J. 2013. Review of: Sandra Mitchell (2009). Unsimple truths. Science, complexity, and policy. *Science & Education* 22: 411–418.
Van Bouwel, J. 2014. Explanatory strategies beyond the individualism/holism debate. In *Rethinking the individualism/holism debate*, ed. J. Zahle and F. Collin. Berlin: Springer.
Van Bouwel, J., and E. Weber. 2002. Remote causes, bad explanations? *Journal for the Theory of Social Behavior* 32: 437–449.
Van Bouwel, J., and E. Weber. 2008. A pragmatic defense of non-relativistic explanatory pluralism in history and social science. *History and Theory* 47: 168–182.
Van Bouwel, J., L. De Vreese, and E. Weber. 2011. Indispensability arguments in favour of reductive explanations. *Journal for General Philosophy of Science* 42(1): 33–46.
Weber, E., and J. Van Bouwel. 2002. Can we dispense with the structural explanations of social facts? *Economics and Philosophy* 18: 259–275.
Weber, E., J. Van Bouwel, and L. De Vreese. 2013. *Scientific explanation*. Dordrecht: Springer.

Shifting Attention from Theory to Practice in Philosophy of Biology

C. Kenneth Waters

1 Introduction

Traditional approaches in philosophy of biology focus attention on biological concepts, explanations, and theories, on evidential support and inter-theoretical relations. Newer approaches shift attention from concepts to conceptual practices, from theories to practices of theorizing, and from theoretical reduction to reductive retooling. They point towards broadening the scope of philosophical attention to investigation, and hence towards analyzing how the integration of practical know-how, concrete knowledge, investigative strategies and theoretical knowledge provides the basis for systematic investigation of the biological world. In this article, I describe the shift from theory-focused to practice-centered philosophy of science and explain how it is leading philosophers to abandon the fundamentalist assumptions associated with traditional approaches in philosophy of science and to embrace scientific pluralism.

This article comes in three parts, each illustrating the shift from theory-focused to practice-centered epistemology. The first illustration concerns conceptual practice in contemporary genetics. I show that geneticists have a flexible concept of the gene that can partition a DNA molecule in a multiplicity of ways. Shifting philosophical attention to conceptual practice reveals how biologists succeed in identifying and manipulating causal strands within systems of bewildering complexity. The second illustration concerns current theorizing about major evolutionary transitions (such as the transition from unicellular to multicellular organisms). This illustration suggests that shifting from the traditional assessment of different theoretical models (in an attempt to identify and articulate the right one) to an analysis of how these models

C.K. Waters (✉)
Minnesota Center for Philosophy of Science & The Department of Philosophy, University of Minnesota, 831 Heller Hall, 55455 Minneapolis, MN, USA
e-mail: ckwaters@umn.edu

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_9,

function in practice provides a more illuminating approach for philosophizing about theoretical knowledge in evolutionary biology. The third illustration concerns reductionism. I show how framing reductionism in terms of the reductive retooling of practice, rather than in terms of theoretical or explanatory relations, offers a more informative perspective for understanding why putting DNA at the center of biological research has been incredibly productive throughout much of biology. Each illustration begins by describing how traditional theory-focused philosophical approaches are laden with fundamentalist assumptions and then proceeds to show that shifting attention to practice undermines these assumptions and motivates a philosophy of scientific pluralism.

2 From Concepts to Conceptual Practices

Traditional analyses of scientific concepts typically presuppose that the most basic concepts of mature sciences designate the fundamental entities and processes of the respective domains (Bird and Tobin 2012). This presupposition is connected to the idea that parsing nature into its fundamental units is the conceptual basis for comprehensively understanding any domain of nature. This philosophical assumption is often reinforced by scientists' own descriptions of their disciplines. For example, geneticists writing textbooks and on-line glossaries have often written that genes are the "fundamental units of heredity". This implies that *the concept of the gene should designate fundamental units in nature, not merely the basic units of a theory*. The assumption that science should be based on concepts that designate fundamental units in nature, which is generally shared by philosophers who have examined the gene concept, is a point I contest. But I also argue that the molecular gene concept, correctly understood, is a powerful concept that provides the basis for a very sophisticated conceptual practice. While philosophers of biology are largely skeptical about molecular gene concepts (e.g. Hull 1974; Burian 1986; Kitcher 1992; Keller 2000), I am skeptical of the philosophical assumptions that underwrite their skepticism.

The idea that the basic concepts of a science should designate fundamental units in nature naturally leads biologists and philosophers of biology to pose the question, "*what is a gene?*" What philosophers have found is that biologists do not have a consistent, coherent, and general answer to this question. The most common answer in the biological literature is that *a gene is a segment of DNA that codes for a protein*. Sometimes this idea is expressed in terms of production: genes are the segments of DNA that produce RNA molecules and RNA molecules are the entities that produce polypeptides. Polypeptide production is important because proteins are comprised of polypeptides, and proteins are very important molecules. They are the building blocks of many biological structures and they play a critical role in the regulation of many important biological processes. The alleged fundamentally of genes is often summarized by the bold assertion that genes direct the development and functioning of organisms by producing proteins. There are a number of problems with such bold

assertions, but I will confine my critical attention here to the comparatively modest premise that genes are segments of DNA that produce polypeptides.

Philosophers have criticized this common definition of gene for a number of reasons. Some reasons concern the use of "code for" and "produce". The use of "code for" is especially problematic. Scientists sometimes build on the notion that genes code for proteins by saying that the "information" for the protein is contained in DNA. Although the term *information* is prevalent in molecular biology and genomics, no one has worked out an account of information sufficient for these sciences (Sarkar 1996; Griffiths 2001; Waters 2000, 2008b). Furthermore, loose talk about information reinforces the pernicious idea that the information of an organism is coded in its DNA.

The idea that genes are DNA segments that produce RNA molecules that produce proteins is a more promising idea, but it has also been criticized by philosophers. Sometimes this idea is criticized on the ground that the production of polypeptides depends on many kinds of molecules, not just DNA and RNA (Oyama et al. 2001).[1] And, of course, this Millian point is correct. Claiming that DNA is *the* cause of RNA or protein production, when in fact many different molecules play causal roles in the production of these molecules, is not philosophically defensible. Nevertheless, as I argue elsewhere, DNA does play a distinctive causal role in the synthesis of RNA and RNA plays a distinctive causal role in the syntheses of polypeptides (Waters 2007).

The distinctive roles of DNA and RNA can be explained briefly as follows. Differences among linear sequences of nucleotides within different DNA segments (i.e. within different genes) largely determine differences among the linear sequences of nucleotides within different RNA molecules. In turn, differences among the linear sequences of nucleotides within different RNA molecules largely determine differences among the linear sequences of amino acids within different polypeptides.[2] So, we can reformulate the gene concept *roughly* as follows:

> *genes* are segments of DNA that causally determine differences among linear sequences within different polypeptides

I call this the *classical molecular gene concept*.

This concept of the gene largely escapes parity objections of the Millian variety (because, roughly speaking, genes are the actual difference makers in RNA and polypeptide synthesis while other molecules are not[3]). Nevertheless, the classical molecular gene concept is susceptible to three other criticisms that philosophers have offered against gene concepts: it seems vague, admits exceptions, and is ambiguous. It seems vague because it is unclear where a gene begins and where

[1] The idea that there is parity among genes and other elements takes several subtle forms in philosophical discussions and I will not analyze them here.

[2] I say "largely" because often in Eukaryotes, differences in other molecules, including splicing agents, also determine actual differences in polypeptides.

[3] See Waters (2007) for a detailed analysis.

it ends. Are regulatory units that precede the so-called "coding region" part of the gene or not? Are introns, the DNA segments that correspond to parts of mRNA molecules spliced out before polypeptide synthesis, part of the gene or not? This concept admits exceptions because many DNA segments determine linear sequences in RNA molecules that do not determine linear sequences in polypeptides. For example, rRNA and tRNA are not "transcribed" into polypeptides. So, genes for rRNA and tRNA pose exceptions to this concept of the gene. Finally, the existence of alternative splicing of mRNA gives rise to ambiguities because one DNA segment contains a multiplicity of overlapping segments that determine linear sequences in a set of different polypeptides that result from differential splicing of the same RNA molecule.

What is a gene? Most philosophers weighing in on this issue have concluded that trying to answer this question is hopeless. Evelyn Fox Keller (2000) suggests that the term "gene" has outlived it usefulness. In apparent frustration, Philip Kitcher (1992) concedes that a gene is anything a competent biologist wants to call a gene. Much of the philosophical literature on this topic implies that the fundamental units of genetics exist at smaller scales (e.g. promoters, enhancers, exons, and introns) or at larger scales (e.g. the level of the gene of classical genetics). Other philosophers have proposed novel gene concepts that seem to depart significantly from conceptual practice. For example, one idea is that genes are processes rather than entities (Griffiths and Neumann-Held 1999). But for the most part, philosophers have decided that the science of today tells us that there is no such thing as a gene at the molecular level.

What about biologists? What do they say when repeatedly pressed by philosophers to answer the question, "What is a gene?"[4] In my own experience, after being shown that their answers are vague, admit exceptions, or are ambiguous, biologists typically shrug their shoulders. Many quickly concede that they do not know *exactly* what a gene is. Reflective biologists add that trying to answer the question with the kind of rigor that philosophers demand would be counterproductive. Progress in genetics, they say, has depended and continues to depend on "muddling through". Science, they insist, would be stymied if geneticists were forced to agree on using a clear and unambiguous concept of the gene.

When biologists say "science would be stymied" and that it is better to "muddle through" they have shifted the topic from theory to practice. This suggests that what philosophers should be analyzing are not fixed concepts, but conceptual practice. We should be examining how geneticists are "muddling through". After all, the practice

[4] A few biologists have written on this issue and drawn conclusions similar to those of philosophers (e.g. see Portin 1993; Fogle 2000). Stotz et al. (2004) have put the question "what is a gene?" to biologists through surveys, keeping track of how biologists in different groups (for example different fields, of different ages, etc.) answer this question. They have explored how biologists answer the question in the context of different kinds of examples. As I read the empirical results, their study indicates that biologists are all over the place. But I have reservations about drawing philosophical conclusions from such studies. See Waters (2004a, b) for a critique of using surveys to analyze scientific concepts.

of genetics has been and continues to be spectacularly successful. Philosophers should analyze the reasoning that makes this practice succeed.

A careful analysis of how contemporary geneticists reason when they use the term "gene" reveals that they use a multiplicity of concepts. Sometimes it is useful to be vague, and in such contexts biologists invoke a blunt concept of the gene from classical genetics, a concept I have called the *classical gene concept* (see Waters 1994b, pp. 165–174, for an analysis of this concept). But in other contexts it is important to be precise. When precision is important biologists employ what I call the *molecular gene concept* (Waters 1994b, 2000).

The molecular gene concept has placeholders. When the placeholders are filled, the concept picks out a precise segment of DNA. So this concept is precise. But it is also flexible because the placeholders can be filled out in a multiplicity of ways, each of which will precisely pick out different segments of DNA. For example when the placeholders are filled out in one way, the concept picks out DNA segments that include introns, when it is filled out in certain other ways it picks out DNA segments that do not include introns. The molecular gene concept is a "gene for" concept and can be articulated as follows:

> **a gene *g* for linear sequence *l* in product *p* synthesized in cellular context *c*** is a potentially replicating nucleotide sequence, n, usually contained in DNA, that determines the linear sequence *l* in product *p* at some stage of DNA expression

The referent of any gene, *g*, is a specific sequence of nucleotides. The exact sequence to which a *g* refers depends how the placeholders *l, p, and c* are filled out. As Fig. 1 illustrates, this provides biologists with the conceptual means to pick out precisely what DNA segments determine different linear sequences in different stages and contexts of DNA expression.

The molecular gene concept is a remarkable conceptual tool. It gives biologists the flexibility they need to identify precise causal threads within incredible complexities of DNA expression. It does so by providing the basis for partitioning the DNA molecule in a multiplicity of ways. There is not one, uniquely correct and comprehensive way to divide DNA into genes. There are lots of ways to divide DNA and the molecular gene concept provides the conceptual means for biologist to do so.

The molecular concept of the gene does not divide DNA at its fundamental joints because the molecule has no such joints. Hence, instead of asking "what is a gene?", with the presumption that genes are units in the uniquely correct *partitioning*[5] of DNA, we should be asking "how should biologists conceive of genes and why?"[6] Answering this question indicates that biologists should conceive of genes in a

[5]I am using the term *partition* in the set theoretical sense of a division into elements that do not overlap.

[6]To be more precise, I believe we should be asking two questions: (1) "what concepts of the gene are at work in successful biological practices?" (2) "what concepts of the gene help us understand the success of biological investigations without inflating the knowledge that makes this success possible?"

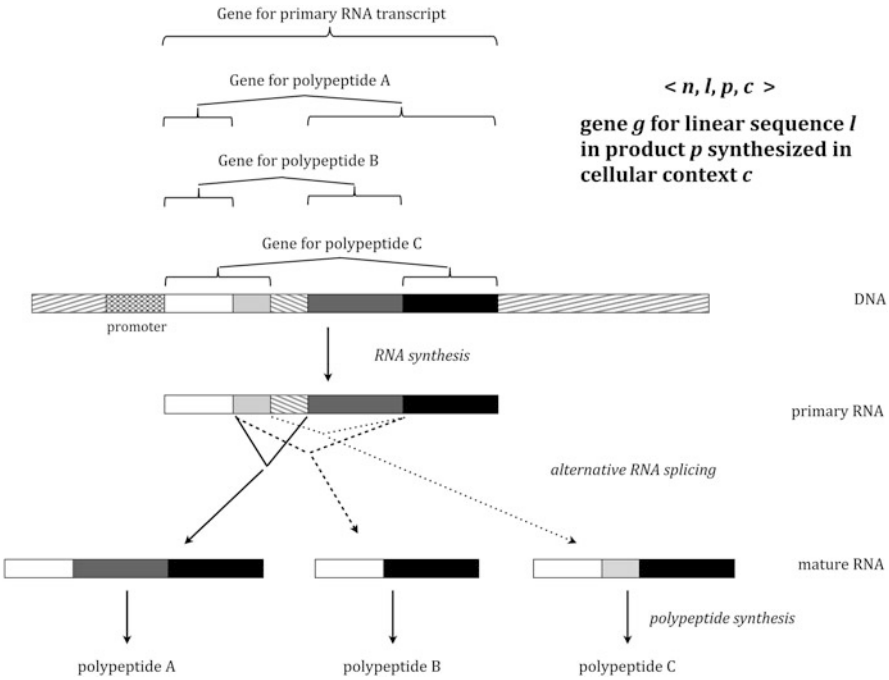

Fig. 1 The molecular gene concept enables biologists to partition DNA in multiple ways. This figure shows how DNA can be partitioned to specify the gene for the primary RNA transcript, and the genes for polypeptides A, B, and C

multiplicity of ways. In some contexts, precision is not important or possible, and biologists conceive of genes in much the same way as classical geneticists did. In other contexts, precision is important. In these contexts biologists should employ the molecular gene concept because it provides a flexible and precise way to identify functional units and this enables biologists to slip and slide through the causal complexities of the biological world.

3 From Theories to Theoretical Practices

Philosophers view scientific knowledge largely through the central theories and explanations of science. Most work in the philosophy of particular sciences involves analyzing, reconstructing, and extending theories and explanations, and examining how these theories and explanations are or could be justified. The traditional approach is to analyze the central theory of a discipline in order to reveal what is essential for in-principle explanation of everything in the discipline's domain. This approach presupposes that the aim of scientific theorizing is to identify the

fundamental relationships (usually presumed to be causal) that are universally responsible for a domain of processes. Philosophical research is often motivated by the assumption that there must be a single right way to formulate the fundamental principles of a mature discipline. The right way to formulate the principles is the one that provides the uniquely correct and *comprehensive* explanation of the discipline's domain. The traditional philosophical goal is to identify and articulate this single right way.

This approach is well illustrated by work in philosophy of evolutionary biology, which has largely centered on analyzing what philosophers take to be the intellectual core of the scientific discipline: Darwin's theory of natural selection, the theory of population genetics, or more recently the theory of evolutionary genetics based on Price's equation. Philosophers have tried to formulate ideal versions of these theories, and along the way have explored philosophical issues, and scientific ones as well. Elliott's Sober's groundbreaking work, *The Nature of Selection* (1984), did just that. Sober reconstructed what he took to be the core of evolutionary biology, a theory of population genetics in which natural selection plays the key role, used his reconstruction to frame and explore a number of philosophical issues including ones concerning causation and evidence. He also employed his analysis to address a controversial issue in evolutionary biology about the levels of selection. The approach exemplified in Sober's work continues to dominate philosophy of evolutionary biology. Although philosophers' interest in the levels of selection issue waned as they took up other scientific controversies (such as ones concerning the importance of natural selection and the connection between evolution and development), philosophical research remains theory-focused. Furthermore, and this is a point I wish to stress in this section, philosophers' analyses of evolutionary theory *often* continues to be framed by the fundamentalist ideal. Although philosophers now disagree about whether the fundamental theory is (or should be) a theory of population genetics, evolutionary genetics based on Price's equation, or a theory that fuses evolution and development, many remain passionately committed to the ideal that there must be one best way to understand evolution, one right way of parsing the causes that divides evolutionary processes into distinct and non-overlapping kinds.

Before proceeding, I should make it clear that I am describing the dominant program of research in the philosophy of evolutionary biology. There are exceptions. Some philosophers are analyzing evolutionary theory from pluralist perspectives (e.g. Waters 1994a, 2005; Sterelny 1996) and some scientists are writing along similar lines (e.g. Dieckmann and Doebeli 2005). Several philosophers, following William Wimsatt's lead (1976, 1980, 1987), have been theory-focused, but have been analyzing practices of theorizing in evolutionary biology without necessarily assuming fundamentalist ideals. In addition, philosophers are beginning to analyze a greater swath of research approaches in evolutionary biology.[7] I should also stress that *there is no necessary connection between focusing on theory and*

[7]For example, a symposium at the most recent Philosophy of Science Association meeting examined experimental modeling in evolutionary biology (Waters et al. 2012).

adopting fundamentalist ideals. Nevertheless, the dominant program of research in the philosophy of evolutionary theory is not only to analyze the products of theoretical practices, but also to analyze them from a monistic perspective with the goal of identifying and articulating the uniquely correct and *comprehensive* basis for parsing and explaining all evolutionary phenomena.[8] This is what I mean when I use the term *traditional* theory-focused approach.

One scientific issue that has recently come to the forefront of philosophers' attention concerns major evolutionary transitions, such as the transition from prokaryotes to eukaryotes and the transition from unicellular organisms to multicellular organisms. How evolution proceeds through such transitions has drawn increasing attention among evolutionary biologists over the past few decades. A central theoretical idea is that these transitions involved groups of organisms evolving into individuals. For example, biologists theorize that multicellular organisms have evolved from groups of conspecific single-celled organisms. How did this evolution proceed? Evolutionary biologists are pursuing a number of different approaches to answer this question including experimental (e.g. Ratcliff et al. 2012), historical/comparative (e.g. Herron and Michod 2007), informal theorizing (e.g. Kirk 2005), and highly abstract mathematical theorizing (see below). Philosophical attention, of course, has gravitated towards analyzing the most abstract theorizing on this issue because this is the approach that presumably gets at the fundamentals.

Much of the abstract theorizing about the evolution of multicellularity centers on the idea that cooperation among single-celled organisms in groups evolved by natural selection until the groups exhibited the characteristics of multicellular individuals. Theoretical biologists are employing abstract models of multilevel selection to represent how this process proceeded. By happy coincidence (perhaps), philosophers have already analyzed multilevel selection theory and reconstructed two different kinds of multilevel selection (Mayo and Gilinsky 1987; Damuth and Heisler 1988). These kinds or model-types have been called *MLS1* and *MLS2* (multilevel selection types 1 and 2). The fact that biologists equivocate between these different models raises interesting philosophical questions for the fundamentalist philosopher. What *is* multilevel selection, is it MLS1 or MLS2? Or maybe multilevel selection comes in two forms? Perhaps some processes of multilevel selection take one form, correctly modeled by MLS1, and other processes of multilevel selection take the other form, correctly modeled by MLS2. But if this is the case, then there must be a more fundamental theory that subsumes both model-types. What is this theory? The philosopher (i.e. the fundamentalist philosopher) wants to know.

These questions have been taken up by Samir Okasha (2006), who has concluded that MLS1 and MLS2 model-types identify different kinds of evolutionary

[8]Consider, for example, the two most recent books in philosophy of biology to win the Lakatos Award, Okasha (2006) and Godfrey-Smith (2009). Both books adopt what I call a fundamentalist perspective.

processes, both of which are subsumed under Price's equation.[9] Okasha's analysis of biologists' theorizing about how evolution proceeds through the transition to multicellularity leads him to conclude that the transition involves successive processes, and that one model-type (MLS1) accounts for early processes (or "stages") in the evolutionary transition and the other model-type (MLS2) accounts for later processes (or "stages"). If we adopt the fundamentalist perspective of traditional theory-focused philosophy of science, as Okasha does, we would assume that these model-types represent distinct kinds of causal processes. On this view, for any particular process of multi-level selection, either an MLS1 model or an MLS2 model, but not both, provides the correct causal account. We might say of one token process, "this *is* MLS1" and of another "this *is* MLS2". We would not, however, say of one process that it is both MLS1 and MLS2.

This analysis, however, raises a difficult question. Given that the evolution from unicellular organisms to multicellular individuals was continuous, how did the evolutionary processes transition from being MLS1 to being MLS2? By leap, or by gradual transition? Were intermediate stages both MLS1 and MLS2? The strict fundamentalist has to say no. MLS1 and MLS2 are distinct kinds. Any single process must belong to at most one of these kinds. Okasha suggests that the evolutionary transition to multicellularity involved intermediate stages in which MLS1 and MLS2 processes were separately occurring (Okasha 2006, p. 59) But he offers no analysis of how this is possible, and no argument why this must be so. It seems to be a conclusion reached on pure faith in the ideal that complex transitions must neatly decompose into processes falling into distinct kinds for which there are in-principle explanations grounded in fundamental principles.

There is another philosophical approach. If we center our attention on the practice of theorizing (instead of trying to interpret, in fundamental terms, the theories that emerge from that practice), what we observe is that evolutionary theorists employ different model-types, and one model-type (MLS1) is better suited to describing causation in the early stages of the evolution of multicellularity, and the other model-type (MLS2) is better suited to describing causation in later stages. Perhaps the middle stages are so messy that neither model-type can be used to describe the causation cleanly. Perhaps, as in the case of molecular biology described in Sect. 1, these evolutionary processes are so incredibly complex that there is no universally applicable parsing of the causes. Why should we believe evolutionary processes must fall neatly into distinctive kinds of processes, that each of these kinds of

[9]Okasha claims that "unlike most formal descriptions of the evolutionary process, it [Price's equation] rests on no contingent biological assumptions, so always holds true" (p. 19). He also claims that the Price formalism "subsumes all more specific models as special cases" (p. 3). But he contradicts this latter claim later in his book, and there is good reason to think that the kind of toolbox theorizing I am advocating with respect to model types MLS1 and MLS2 applies at the level of the Price equation and its formal rivals such as contextual analysis (see Waters 2010).

processes can be fully represented by one model-type, and that the collection of model-types explaining the different kinds of processes can be subsumed under a single set of fundamental principles?

If evolutionary processes do not fall into distinctive kinds, then theorizing as if they did, might be counterproductive in the same way as insisting upon a single rigid concept of gene would be counter-productive in genetics. Theoretical evolutionary biologists need to slip and slide through the complexities of evolutionary history just as molecular biologists need to slip and slide through the complexities in their domains of investigation. The philosophical task, I submit, is to analyze how theoreticians succeed to make sense out of the complexities.

I began this section by describing the fundamentalist view embedded in the traditional approach to analyzing scientific theories. Of course, philosophy, like biology, is pluralistic and there is no necessary connection between focusing on theoretical products and holding the fundamentalist view.[10] For example, philosophers of biology interested in systematics, which is an important part of evolutionary biology, have developed accounts of natural kinds that are more suited to represent the blurring and merging of kinds (e.g. Dupré 1981; Ereshefsky 1998). But it is unclear how accounts, such as Boyd's homeostatic cluster account of natural kinds (1999), could be applied to processes. In any case, when philosophers analyze the basic theories of evolution, the analyses typically presuppose old-fashioned ideas about natural kinds *of processes*, fundamental principles, and universal in-principle explanation.

It is appropriate to conclude this section by drawing an explicit contrast between the perspective of theorizing presupposed by the traditional theory-focused approach and the view that emerges from centering philosophical attention on the practice of theorizing.

> **Fundamentalist view** presupposed by the traditional analysis of theories: the aim of scientific theorizing is to identify the fundamental causal relationships that are universally responsible for a domain of processes. Achieving this aim entails articulating the fundamental theoretical concepts and causal principles that can provide a basis for constructing models that decompose the causes of each and every process in the uniquely correct way. Proponents of this view stress the idea that there is, of course, just one way the world actually is, and the aim of theorizing is to describe, in a principled manner, the one way it actually is.

> **Toolbox view**[11] that emerges from centering philosophical attention on practices of theorizing: one aim of scientific theorizing is to construct causal models that explain aspects of the processes in a domain. Achieving this aim entails articulating a multiplicity of theoretical concepts and causal principles that can be drawn upon to construct models that might decompose the causes of different processes in different ways and the causes of some

[10] I thank Marc Ereshefsky for reminding me that theory-focused philosophers of biology have done a good job critiquing the fundamentalist conception of natural kinds and that they have developed promising alternatives for understanding kinds of entities.

[11] Maxwell (manuscript), Cartwright et al. (1995), Cartwright (1999), Suárez and Cartwright (2008) and Wimsatt (2007) offer ideas about theorizing similar to the one I am advancing here and also use the "toolbox" term and metaphor.

> processes in a multiplicity of ways. In cases of multiplicity, some concepts and models offer the best account of some aspects of a given process, other concepts and models provide the best account of other aspects.

The toolbox view does not deny that the world is one way. But it rejects the assumption that there must be one best way to describe the world, and that this one best way offers a principled, unified, and comprehensive basis for explaining every aspect of the world. On the toolbox view, part of the work of philosophy is to formulate ideal accounts of how scientists should theorize in such a world. To achieve this, philosophers should analyze practices of scientific theorizing, especially those involving complex phenomena such as evolution.

4 From Theory Reduction to Reductive Retooling of Practices

Accounts of reductionism in philosophy of science focus on theoretical and explanatory relations. The basic idea, which was set out in canonical form by Ernest Nagel (1961), is that the core principles of a reduced theory are derived (or explained) by core principles of a reducing theory.[12] Nagel combined this formal idea with a historical one: scientists first establish a "higher-level" theory, and afterwards reduce it by deriving its principles from a "lower-level" theory. The higher-level theory is revised in the process of being reduced, and the revision provides more accurate explanations and predictions over a greater range of phenomena. Hence, theoretical reduction purportedly advances science towards integrated, monistic, and comprehensive knowledge.

Kenneth Schaffner (1969) applied Nagel's model to biology and claimed that the theory of classical genetics was being reduced to a theory of molecular genetics. Schaffner's claim has been roundly rejected by philosophers of biology (Hull 1974; Wimsatt 1976; Darden and Maul 1977; Hooker 1981; Kitcher 1984; Rosenberg 1985). Yet few critics would now deny that the discipline of genetics has in some sense gone molecular and that the molecularization of genetics has led to a dramatic transformation of much of biology. Questions about how the molecularization of genetics transformed the biological sciences, however, have fallen through the cracks of philosophical inquiry.

Critics of Schaffner's account have typically abandoned reductionism, but not his focus on theories and theoretical explanations. From the traditional theory-focused

[12] Some recent accounts of reduction frame reduction in different ways, but still with the emphasis on theoretical and/or explanatory relations. For example, Hüttemann and Love (2011) couch it in terms of explanations in which an outcome described at a higher level (explandum) is explained by earlier states described at lower level(s). This article illustrates that focusing on the theories and explanations does not necessarily presuppose fundamentalism, and that paying attention to theoretical and explanatory practices undermines the fundamentalist ideals.

perspective, the lack of *impressive* reductive explanations (or derivations) of the central principles of classical genetics suggests that there is nothing more to say about the reduction of genetics.[13] But if we stop focusing exclusively on theoretical developments and broaden our attention to the investigative practice of classical genetics, an impressive reductive retooling of this practice becomes visible. The reductive retooling of classical genetics is the source of the dramatic transformation of much of biology into DNA-centered science.[14] In many biological sciences, it is the investigative strategies and procedures, not the explanations, that center on DNA.

Traditional conceptions of theoretical reduction assume that the aim of science is to provide a unified and comprehensive, theoretical explanation of the world, an explanation based on fundamental principles (or laws). This view, as exemplified in Nagel's model of theoretical reduction, takes the structure of scientific knowledge to resemble the structure of a layer cake. Each layer of science (physics, chemistry, biochemistry, genetics, ...) searches for the principles that are essential for explaining everything at the level of its domain. Reductionists claim that the principles at higher levels of the cake are "reducible" to principles at lower levels. In Nagel's version of reductionism, this means that the principles at higher levels are derivable from principles at lower levels. Antireductionists generally accept the basic layer cake picture of scientific knowledge. But they disagree with reductionists about how successive layers of the cake relate to one another. Antireductionists argue that the principles essential for explaining domains of phenomena at higher levels are not reducible to lower-level principles.[15]

Reductionists and antireductionists also share a common epistemic ideal about knowledge within each layer of scientific knowledge.[16] According to this ideal, scientists seek to discover and establish the uniquely correct theory that provides an in-principle explanation of everything in a domain. Investigation in mature sciences is successful according to this ideal because the core theories get *the fundamentals* basically correct and this provides the basis for explaining a greater range of phenomena with increasing accuracy.

The theory-focused view of theoretical reductionism can be summarized as follows. Reductionism contributes to the advance of a science by connecting theories at higher levels, which have relatively narrow domains, to theories at lower levels, whose domains are broader. In doing so, the reduced theories are improved

[13]Kitcher (1984) offers an alternative, theory-focused, non-reductive account of how molecular genetics contributes to the science of genetics. But Hull (1974) implied that there was something reductive happening in genetics, but that what was happening did not fit Nagel's model.

[14]Many of the points in this section are developed in more detail in Waters (2008a).

[15]Not all reductionists accept the layer-cake image (e.g., Weber 2005), and some antireductionists seem more interested in advancing holism than multileveled holism (e.g., various contributors to Oyama et al. 2001). Nevertheless, many philosophers cling to the idea that biology is organized into separate sciences, each of which is focused on a particular level of organization.

[16]This epistemology is generally presupposed even by reductionists and antireductionists who reject the layer-cake image (e.g. Weber 2005; Oyama et al. 2001).

upon and provide better bases for explaining a greater range of phenomena within their domains. This view offers an account of why reduction is fruitful, the feature of reduction which Nagel himself claimed was most important. Reduction is fruitful, on this account, because it improves theory, and improving theory is what inquiry is all about!

Practice-centered epistemology leads to a different account of why reductionism is fruitful. The practice-centered view does not assume that science has any single aim such as improving theory.[17] Science is practiced by scientists with a multiplicity of aims. Among the commonly shared aims are the quests to investigate, manipulate, and explain phenomena. In mature sciences, such as classical genetics (by the mid1920s), scientists have the means to investigate systematically a domain of phenomena. But this is not necessarily because they have a core theory that provides an in-principle explanation of everything in the domain. That is, the success of their practice does not rest on a core theory that grasps the *fundamentals*.

The core theories of sciences like classical genetics, explain very little, even in-principle, with respect to the domains being systematically investigated. But in addition to core theories, scientists have concrete knowledge, procedural know-how, and strategic approaches that enable them to investigate ranges of phenomena that far outstrip the ranges of phenomena that their core theories can possibly explain. What provides the bases for systematic investigation are not the core theories by themselves, but the *investigative matrices* into which these theories are assimilated.[18]

In the case of classical genetics, the core theory was the transmission theory of genetics. All it could explain, in-principle, was the transmission of phenotypic differences from one generation to the next. But it played an important role in the strategic approach that classical geneticists used to investigate a wide variety of basic processes. Their approach, "*the genetic approach*", was to (a) identify naturally occurring or artificially produced mutants that exhibit a difference relevant to some biological process of interest, (b) carry out genetic analyses of the mutants, and (c) recombine the mutants to learn more about the process of interest. Theory was integral to carrying out each of these steps, but doing so was also depended on concrete knowledge, procedural know-how, and strategic approaches. What is critical for understanding the power of genetics is what has been completely overlooked in the uniformly theory-focused accounts of classical genetics contained in the philosophical literature. Namely, the *core strategy* of the genetic approach: classical geneticists sought to use mutations as tools to disrupt the processes they wished to investigate, and in disrupting the processes they sought to learn about how the processes worked.

[17]Debates in philosophy of science are sometimes framed as disagreements about "the aim" of science. For example, van Fraassen (1980) characterizes his disagreement with scientific realists as centering on the aim of science. I reject the idea that there is something called "science" that has a single aim.

[18]See Waters (2004b) for an elaboration of this account.

The basic strategy of learning about biological processes by disrupting them wasn't new. For centuries, physiologists investigated mechanisms, such as the mammalian circulatory system, by interfering with its parts and observing what happens when the processes are subsequently disrupted. In fact, amateur mechanics use this investigative approach to investigate how machines work. What was new to classical genetics was the idea that one could disrupt processes, such as sex determination, by recombining genetic mutations.

The genetic approach had mixed success in classical genetics. It tended to be most successful in the study of chromosomal mechanics. But the approach was not confined to investigating chromosomal processes. It was also used to learn about basic biological processes such as gene action, mutation, development, and evolution. For example, mutants involving dosage effects and genetic mosaics were often investigated in order to shed light on gene action or on broader issues of development, such as sex determination. Generally, this research did not live up to the promise of yielding important new knowledge about the basic processes being investigated (at least in the short run). But as Robert Kohler (1994) observed, even when experiments did not lead to new knowledge about the phenomena being investigated (such as sex determination), they still yielded publishable information about the existence and genetic location of new alleles. This helps explain why the underlying investigative strategy that systematized research in classical has subsequently disappeared from view, and why it appears from hindsight that the point of the experimentation was to discover and map new genes.

The epistemology of scientific practice exemplified by classical genetics can be summarized as follows. Investigative practice has two domains. One range, the *explanatory range of its core theory*, is the domain of what can be explained, at least in principle, by its core theory. The other, which I call the *investigative reach*, is what can be systematically investigated and explained in piecemeal fashion by employing the *investigative matrix*, which consists of an assimilation of core theory, concrete knowledge, procedural know-how, and strategic approaches. Classical genetics illustrates how the explanatory range of a core theory is much narrower than the investigative reach of the investigative practice. The core theory in this case, the transmission theory, could explain only the transmission of phenotypic differences from one generation to the next. But the genetic approach drew upon this theory, concrete knowledge, and procedural know-how in an effort to systematically investigate a broad range of basic biological processes. If successful, these efforts yield piecemeal explanations of parts and aspects of processes outside the explanatory range of the core theory. In the case of classical genetics, however, these efforts rarely yielded such explanations.

With this practice-centered understanding of classical genetics in place, we can answer questions about what the molecularization of genetics consisted of, how it turned genetics into such powerful science, and how genetics subsequently transformed much of biology. In brief, molecularization of genetics consisted of a retooling of investigative practice. When mutations were identified with physical differences in DNA, they could be screened, isolated, moved, and tracked with remarkable precision. Geneticists learned how to engineer new mutations. These

procedural gains make it possible to deftly manipulate all kinds of biological processes, and this has greatly enhanced the power of the genetic approach of investigation. This power is being wielded by investigators throughout the biological sciences. This is what has transformed biology.

A typical example of DNA-centered research in neuroscience illustrates how the genetic approach systematizes research without being guided by a core theory (i.e. without being guided by a theory that alleged offers an in-principle explanation of the process being investigated). Among the processes investigated in neuroscience is the formation of neurological systems during the development of individual organisms. Investigation often takes the form of identifying the functional role of various elements in particular processes or mechanisms of development.

For example, Bastiani, Jorgensen, and Hammarlund at the University of Utah have investigated the establishment of neural connections between the ventral and dorsal nerve chords in the nematode, *C. elegans* (Hammarlund et al. 2007).[19] Investigators had already discovered that a protein, ß-spectrin, is located in the growth cones of nerve cells that are growing from the ventral to dorsal nerve chords. This discovery led to the hypothesis that ß-spectrin plays a functional role in the growth of these neurons (as they are forming the connection between ventral and dorsal nerve chords). But Bastiani, Jorgensen, and Hammarlund learned that the role of ß-spectrin is to protect the delicate neuronal structure that connects the nerve chords from acute strains. That is, they learned, that ß-spectrin serves its role after the connection is made, and apparently not before. They learned this by using the genetic approach.

Bastiani, Jorgensen, and Hammarlund disrupted the developmental processes they were investigating by manipulating genes. They made the processes visible by inserting a gene for green fluorescent protein (GFP). Obviously this gene did not play a role in the explanation of neuronal growth or maintenance. In the first stage of their investigation, the researchers prevented ß-spectrin from being synthesized in experimental organisms by using a mutated version of the gene for ß-spectrin. They learned that the neurons developed normally, which suggested that the role of ß-spectrin is to prevent degeneration of the neuronal extensions, not facilitate their growth.[20]

In the next stage of experimentation, the investigators immobilized worms by interfering with the expression of the gene for myosin. They observed that the neurons of immobilized worms lacking ß-spectrin maintained their structure about as well as worms with ß-spectrin. These results were checked by manipulating a fourth gene, twitchen, which did not immobilize the worms but did prevent them from moving in ways that would exert acute strains on the relevant nerve cells

[19]As in the case of classical genetics (see Waters 2004b), investigators carry out their work on model organisms that have been adapted for laboratory practice.

[20]Given the possibility of redundant pathways to the development of the neurons, the results did not prove that ß-spectrin has no role in the growth if these neuronal extensions, but the results did show that the role was not essential.

extensions. They observed that the neuronal connections in twitchen worms lacking ß-spectrin were maintained about as well as the neurons as worms with ß-spectrin. The combination of results indicated that the functional role of ß-spectrin is to protect neurons against acute strains.

Bastiani, Jorgensen, and Hammarlund's explanation about the function of ß-spectrin is appropriately couched in terms of proteins and cytoskeletal structures, not in terms of genes or DNA. These neuroscientists intervened on genes and DNA expression to manipulate the processes they were investigating, but genes and DNA were not part of the mechanistic explanation that emerged from their research. The point worth emphasizing is that their research was not guided by an attempt to fill out the details of a gene-centered theory. This example illustrates how investigation can be systematized by the genetic approach, an investigative strategy, without being guided by a comprehensive gene-centered theory about the phenomena to be explained.

The difference between the conception of reduction that emerges by centering attention on investigative practice instead of focusing on theory and explanation can be summarized with three contrasts. First, whereas the theory-focused conception of reduction holds that core principles of the reduced theory are derived (or explained) by core principles of a reducing theory, the practice-centered conception holds that elements of an investigative practice are associated with elements at smaller scales. For example, in the retooling of genetics, mutations in genes on chromosomes are associated with differences in linear sequences in DNA (see Weber 2005, pp. 157–169). Second, whereas focusing on theory assumes the power of reduction must be an increase in explanatory power of a reduced theory, centering attention on practice reveals that the power of reduction stems from an increase in the power to manipulate and hence investigate a greater range of processes. Retooling increased the power of genetics because it enabled biologists to manipulate and track a wide range of processes by intervening on DNA. Third, embedded within the theory-focused view of reduction is the assumption that the role of reduction is to advance science towards an integrated, monistic, and comprehensive understanding of the world. According to the practice-centered view, reductive retooling advances science towards more powerful means of control and a fragmentary understanding of a greater number of aspects and parts of the world.

5 Conclusion

Philosophy of science is, as Ian Hacking (1983) remarked, theory-biased. The bias often involves not just focusing attention on theories and explanations, but also on using an approach that presupposes a metaphysics and epistemology of fundamentalism. Such bias leads to a distorted understanding of scientific knowledge and practice, one that attributes the success of a mature science to its core theory providing an in-principle, comprehensive explanation of everything in the science's domain.

Theory-bias obscures the most remarkable feature of science: scientists' ability to manipulate, explain, and predict aspects of and parts of a world when they lack the kind of understanding that many philosophers assume they must have (or approximate) in order to be successful. What scientific success actually consists of, how scientists achieve this success, and how we should interpret the results of their success are what I take to be central questions for philosophy of science. The best way to approach these questions is to broaden our attention to the practice of science and to free ourselves from the fundamentalist assumptions that have restricted traditional philosophical inquiry.

References

Bird, A., and E. Tobin. 2012. Natural kinds. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta (Winter 2012 Edition). http://plato.stanford.edu/archives/win2012/entries/natural-kinds/.

Boyd, R. 1999. Homeostasis, species, and higher taxa. In *Species: New interdisciplinary essays*, ed. R. Wilson, 141–186. Cambridge, MA: The MIT Press.

Burian, R.M. 1986. On conceptual change in biology: The case of the gene. In *Evolution at a crossroads*, ed. D.J. Depew and B.H. Weber, 21–42. Cambridge, MA: The MIT Press.

Cartwright, N. 1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cartwright, N., T. Shomar, and M. Suárez. 1995. The tool box of science. In *Theories and models in scientific processes*, ed. W. Herfel, W. Krajewski, I. Niiniluoto, and R. Wojcicki, 137–149. Amsterdam: Rodopi.

Damuth, J., and I.L. Heisler. 1988. Alternative formulations of multi-level selection. *Biology and Philosophy* 3: 407–430.

Darden, L., and N. Maul. 1977. Unifying science without reduction. *Philosophy of Science* 8: 43–71.

Dieckmann, U., and M. Doebeli. 2005. Pluralism in evolutionary theory. *Journal of Evolutionary Biology* 18(5): 1209–1213.

Dupré, J. 1981. Natural kinds and biological taxa. *The Philosophical Review* 90: 66–90.

Ereshefsky, M. 1998. Species pluralism and anti-realism. *Philosophy of Science* 65: 103–120.

Fogle, T. 2000. The dissolution of protein coding genes in molecular biology. In *The concept of the gene in development and evolution. Historical and epistemological perspectives*, ed. P. Beurton, R. Falk, and H.-J. Rheinberger, 3–25. Cambridge: Cambridge University Press.

Godfrey-Smith, P. 2009. *Darwinian populations and natural selection*. Oxford: Oxford University Press.

Griffiths, P.E. 2001. Genetic information: A metaphor in search of a theory. *Philosophy of Science* 68(3): 394–412.

Griffiths, P.E., and E.M. Neumann-Held. 1999. The many faces of the gene. *BioScience* 49(8): 656–662.

Hacking, I. 1983. *Representing and intervening*. Cambridge: Cambridge University Press.

Hammarlund, M., E.M. Jorgensen, and M.J. Bastiani. 2007. Axons break in animals lacking ß-spectrin. *Journal of Cell Biology* 176(3): 269–275.

Herron, M.D., and R.E. Michod. 2007. Evolution of complexity in the volvocine algae: Transitions in individuality through Darwin's eye. *Evolution* 62(2): 436–451.

Hooker, C.A. 1981. Towards a general theory of reduction. Part I: Historical and scientific setting, Part II: Identity in reduction, Part III: Cross-categorical reduction. *Dialogue* 20: 38–59, 201–236, 496–521.

Hull, D.L. 1974. *The philosophy of biological science*. Englewood Cliffs: Prentice-Hall.
Hüttemann, A., and A.C. Love. 2011. Aspects of reductive explanation in biological science: Intrinsicality, fundamentality, and temporality. *The British Journal for Philosophy of Science* 62: 519–549.
Keller, E.F. 2000. *Century of the gene*. Cambridge, MA: Harvard University Press.
Kirk, D.L. 2005. A twelve-step program for evolving multicellularity and a division of labor. *BioEssays* 27: 299–310.
Kitcher, P.S. 1984. 1953 and all that: A tale of two sciences. *Philosophical Review* 43: 335–371.
Kitcher, P.S. 1992. Gene: Current usages. In *Keywords in evolutionary biology*, ed. E. Keller and L. Lloyd, 128–131. Cambridge, MA: Harvard University Press.
Kohler, R.E. 1994. *Lords of the fly: Drosophila genetics and the experimental life*. Chicago: University of Chicago Press.
Maxwell, G. manuscript. Toolbox theorizing. Maxwell Archive, Minnesota Center for Philosophy of Science, University of Minnesota, Minneapolis.
Mayo, D.G., and N.L. Gilinsky. 1987. Models of group selection. *Philosophy of Science* 54(4): 515–538.
Nagel, E. 1961. *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt, Brace & World.
Okasha, S. 2006. *Evolution and the levels of selection*. Oxford: Oxford University Press.
Oyama, S., P.E. Griffiths, and R.D. Gray. 2001. Introduction: What is developmental systems theory? In *Cycles of contingency*, ed. S. Oyama, P.E. Griffiths, and R.D. Gray, 1–12. Cambridge, MA: Bradford/The MIT Press.
Portin, P. 1993. The concept of the gene: Short history and present status. *The Quarterly Review of Biology* 68: 173–223.
Ratcliff, W.C., R.F. Denison, M. Borrello, and M. Travisano. 2012. Experimental evolution of multicellularity. *Proceedings of the National Academy of Science* 109(5): 1595–1600.
Rosenberg, A. 1985. *The structure of biological science*. Cambridge: Cambridge University Press.
Sarkar, S. 1996. Biological information: A skeptical look at some central dogmas of molecular biology. In *The philosophy and history of molecular biology: New perspectives*, ed. S. Sarkar, 187–231. Dordrecht: Kluwer.
Schaffner, K. 1969. The Watson-Crick model and reductionism. *The British Journal for the Philosophy of Science* 20: 235–248.
Sober, E. 1984. *The nature of selection: Evolutionary theory in focus*. Cambridge, MA: Bradford/The MIT Press.
Sterelny, K. 1996. Explanatory pluralism in evolutionary biology. *Biology and Philosophy* 11(2): 193–214.
Stotz, K., P.E. Griffiths, and R.D. Knight. 2004. How scientists conceptualise genes: An empirical study. *Studies in History and Philosophy of Biological and Biomedical Sciences* 35(4): 647–657.
Suárez, M., and N. Cartwright. 2008. Theories: Tools versus models. *Studies in History and Philosophy of Modern Physics* 39: 62–81.
van Fraassen, B. 1980. *The scientific image*. Oxford: Oxford University Press.
Waters, C.K. 1994a. Tempered realism about the force of selection. *Philosophy of Science* 58: 553–573.
Waters, C.K. 1994b. Genes made molecular. *Philosophy of Science* 61: 163–185.
Waters, C.K. 2000. Molecules made biological. *Revue Internationale de Philosophie* 54(214): 539–564.
Waters, C.K. 2004a. What concept analysis should be. *History and Philosophy of the Life Science* 26: 29–58.
Waters, C.K. 2004b. What was classical genetics? *Studies in History and Philosophy of Science* 35: 783–809.
Waters, C.K. 2005. Why genic and multilevel selection theories are here to stay. *Philosophy of Science* 72(2): 311–333.
Waters, C.K. 2007. Causes that make a difference. *The Journal of Philosophy* CIV(11): 551–579.

Waters, C.K. 2008a. Beyond theoretical reduction and layer-cake antireduction: How DNA retooled genetics and transformed biological practice. In *Oxford handbook to the philosophy of biology*, ed. M. Ruse, 238–262. New York: Oxford University Press.

Waters, C.K. 2008b. Molecular genetics. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta (Fall 2013 Edition), forthcoming http://plato.stanford.edu/archives/fall2013/entries/molecular-genetics/.

Waters, C.K. 2010. Okasha's unintended argument for toolbox theorizing. *Philosophy and Phenomenological Research* 82(1): 232–240 [pre-print of unabridged version on PhilSci-Archive.pitt.edu].

Waters, C.K., K. Hillesland, M. Weber, and M. Travisano. 2012. Can experimental modeling play the role of theorizing in evolutionary biology? *Philosophy of science association 2012 biennial meeting*, San Diego.

Weber, M. 2005. *Philosophy of experimental biology*. Cambridge: Cambridge University Press.

Wimsatt, W. 1976. Reductive explanation: A functional account. In *PSA 1974 proceedings of the 1974 biennial meeting Philosophy of Science Association*, Boston studies in the philosophy of science, vol. 32, ed. R.S. Cohen, C.A. Hooker, A.C. Michalos, and J.W. Van Evra, 671–710. Dordrecht: Reidel.

Wimsatt, W. 1980. Reductionistic research strategies and their biases in the units of selection controversy. In *Scientific discovery: Case studies*, Boston studies in the philosophy of science, vol. 60, ed. T. Nickles, 213–259. Dordrecht: Reidel.

Wimsatt, W. 1987. False models as means to truer theories. In *Neutral models in biology*, ed. M. Nitecki and A. Hoffman, 23–55. London: Oxford University Press.

Wimsatt, W. 2007. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Living Instruments and Theoretical Terms: Xenografts as Measurements in Cancer Research

Pierre-Luc Germain

1 Introduction

As I argue in the first part of this paper, so-called xenograft "models" of cancer are often used not as models in the traditional, analogical sense, but as measuring devices. This prompts the question of what it is that they measure, and of the relationship they entertain with it. To investigate these issues, I compare two cases of xenograft as measurements with the prototypical example of a measuring device: the thermometer. I rely on the work by Hasok Chang on the history and epistemology of thermometry (Chang 2004). Behind the apparent simplicity of thermometers lies a daunting epistemological problem, which he labels "the problem of nomic measurement": in a nutshell, there are a variety of thermometers giving inconsistent (not linearly correlated) readings, and we would need to know already what temperature is in order to know which one gives the right reading. I highlight some relevant similarities between his history of thermometry and the examples I will present from cancer research. In both cases, instruments and theories have a reciprocal stabilizing role: the instruments are at the same time means subordinated to theoretical understanding, and theoretical terms are means of bridging different instrumental and operational contexts. Finally, the comparison sheds some light on a contemporary debate in cancer research.

In the first part of this paper, I present the instrumental role that organisms sometimes play in biomedical research (Sect. 2.1), and apply this concept to early xenograft models of cancer (Sect. 2.2). I show how, in early xenograft experiments, transplantability was taken as a signal for an abstract quantity. Ultimately, this attempt at the mutual stabilization of operational and theoretical concepts failed.

P.-L. Germain (✉)
University of Milan and European Institute of Oncology (IEO), Campus IFOM-IEO, Via Adamello 16, 20139 Milan, Italy
e-mail: pierre.germain@ieo.eu

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_10,

However, this failure may be informative for a similar attempt in contemporary cancer research. In order to better understand the epistemological issues involved, I review how the analogous problem was solved in the case of thermometry (Sect. 3.1). I then present the more recent use of xenograft as measurements in the Cancer Stem Cell Framework, and the specific problem of establishing an operational definition (Sect. 3.2). Finally, I show how some insights from the thermometry example can be used to inform, and even take position on, some of the issues relating to this problem (Sect. 3.3).

2 Living Instruments

2.1 The Roles of Laboratory Animals in Biomedical Research

Animals have long been used in biological research aimed at learning about human biology, most often acting as "a surrogate for a human being" (ILAR and NRC 1998, p. 10). It is this surrogacy that has warranted calling them "models". However in the life sciences there seems to have been a conflation between this intensional meaning of the term and its coincidental extensional meaning, so that organisms *often* used as models came to permanently bear that label. While they are indeed often used as surrogates, this obscures a great variety of functions that organisms actually play in research. For instance, organisms are often factories for materials, as was the case in the first half of the twentieth century when stocks of viruses were kept and grown in the lab by serial infection of host animals (e.g. rabbits). Nowadays, plasmids are routinely used for DNA cloning, and labs around the world still rely on the bleeding of mice (or rabbits, goats, etc.) for the production of antibodies. Calling these animals "models" would be so far-fetched as to rob the notion of any meaning.

Organisms can therefore have a variety of roles in biomedical research, which are not exhausted by the notion of model. The examples that I will discuss here represent a particular such function, which is that of measuring/detection device (which I will call the "instrumental" role for reasons of convenience).

What I mean by instrumental role can be illustrated with a simple example: the Ascheim-Zondek (A-Z) test for pregnancy invented in the 1920s. In this test, mice are injected with the urine of a female patient, and are dissected after 2 days. If the injection caused small blood stains on the mouse's ovarian follicles, then the patient is pregnant (Zondek 1928). The mouse, here, is not used as a replica of the patient, not least because the phenotype actually being used by the test – the blood stains – are absent from the woman. Rather, it seems justified to talk of a measuring (or at least detection) device, as the animal has the function of detecting a signal in order to learn something about the woman. The mouse allows one to detect, in the input, something that was otherwise unobservable.

One can distinguish several kinds of observational instruments, and in order to talk of a measuring/detection device, there needs to be a kind of decoupling between

the observable output of the device (the signal) and what this allows us to infer in the target system: not all that is observable in the readout of an instrument is actually informative about the input. The fact that there is, for instance, an air bubble in the column of mercury is known to be irrelevant to the temperature one wishes to measure. Finally, the case of the A-Z test is not a measuring device in a strict sense, for its output is binary – it is a detection device. A measuring device should at least provide ordinal, if not quantitative, readouts. This implies that "measuring locates the target in a theoretically constructed logical space" (van Fraassen 2008, p. 2), and indeed I will be concerned here with this relationship between measuring devices on the one hand, and the structure and reference points of this theoretical space on the other.

Elsewhere (Germain forthcoming), I characterize the instrumental role of organisms in detail, and argue for its relevance in contemporary research. Here, I would like to pursue a slightly different goal, namely to push the analogy to a comparison with the classical example of the thermometer, in order to study the relation between these instruments and what it is that they should measure. Throughout this paper, I will discuss cases of such "living instruments" which are strikingly similar to the A-Z test, and yet still of high relevance in contemporary cancer research: xenografts as measurements of tumorigenicity.

2.2 Xenograft Experiments in Early Cancer Research

Following the nomenclature of Snell (1964), a xenograft – or xenotransplantation – is a case of tissue transplantation where the donor and recipient are of two different species. From the end of the nineteenth to the middle of the twentieth centuries, transplantation was of big interest to the scientific community, and scientists attempted a disarraying diversity of transplant experiments. This was especially common in the field of cancer, in an attempt to domesticate tumours to the laboratory. Human tumours, if they were to be studied experimentally, needed to be studied outside their host. Even in the case of animal tumours, scientists were confronted with the simultaneous shortage of spontaneous tumours and inability to sustain a tumour beyond the death of its host. Hence transplantation became (and still is today, although for different reasons) among the most widespread ways of studying cancer in a lab.

There is some disagreement as to the first author to be credited with successful tumour transplantation. Claims go back at least to 1889–1898 (Hanau 1889; Mayet 1902; Ewing 1919),[1] but were all strongly criticized – see for instance Hekzog (1902), who instead credited the feat to Loeb (see also Loeb 1945). More recently, a

[1] As a matter of fact, Novinski (1876) showed even earlier the successful transplantation of the canine venereal tumour. However, because it was believed that a virus was transmitted, rather than the tumour itself, Novinsky's work was never interpreted as transplantation. It was only established recently that the tumour itself, and not some infectious agent, is transmitted (Murgia et al. 2006).

historical review of chemotherapy attributes the "first transplantable tumor systems in rodents" to Clowes in the early 1910s (DeVita and Chu 2008, p. 8643). The contention seems to hinge on what is stable enough to constitute a "system". Indeed, an important reason for the disagreement is that for a long time, the criteria on which to evaluate a successful graft were unclear (see Loeb 1945, Chap. 12). In general, grafts lasted only for some time before resorbing under the pressure of the host's immune system (although, at that time, the explanation was that the foreign cells lacked specific "foods" – it was not until Medawar's work in the 1940s that the immunological basis of rejection was firmly established). Therefore, a line had to be drawn somewhere to distinguish cells that have successfully engrafted, albeit only temporarily, and cells that are just "still there" from the injection. Some authors were already discussing histological criteria, for instance vascularization, but for a long time there was no established way to make the distinction. A related problem is that the injection caused an injury to the recipient that had important risks of infections, which (either because of the inflammation or of the death it brought) could easily pass for cancer.

The systematic, large-scale work of Loeb (especially from 1901 to 1910) was certainly of central importance in the establishment of transplantation systems (Witkowski 1983), but the experiments were of limited success for a long time (see for instance Funk 1915). The main improvement in this respect came from the discovery that some locations in the host (the brain, the anterior chamber of the eye, etc.) accepted grafts more readily. As grafts started to become more efficient, and transplantation systems were tamed, the possibility appeared of using them as tools for a variety of purposes.

The long established observation that only embryonic and cancer tissues were transplantable across species (normal adult tissues "did not take") lead to such instrumental uses. Some scientists proposed that "transplantability constitute[s] a biological test for cancer" (Greene 1948, p. 1364). Greene suggested that the "study of the transplants allows a more precise classification than is warranted from the morphologic features of the biopsy specimen" (*ibid.*). He was explicitly proposing a diagnostic tool to replace what he considered to be a "coarse" and uninformed judgement of pathologists.

Importantly, transplantation was not simply believed to be a useful signal: if it was a good signal, it was because it was signaling *something*, and therefore giving access to some invisible differences between cancer cells:

> The fact that a biological quality as fundamental as the ability to grow in an alien species differentiates morphologically identical tumors suggests that the tumors must also differ in metabolic or biochemical constitution. It would seem important, therefore, to distinguish tumors with respect to this property and to study the different groups formed rather than to consider morphological similarity a proof of constitutional identity. (Greene 1952, p. 41)

The very idea of using transplantation as a test implied that transplantation made visible a difference that was already in the tissues. More importantly, transplantability was not understood as binary: degrees of transplantability could be obtained either by resorting to statistics (the proportion of cases where the transplant was

successful) or by assessing the pace, duration, and quality of the growth. Hence more than a tool to detect malignancy, transplantation was a tool to measure it. Arguably, this quantity was not numerical in a strong sense, but it was at least ordinal: by 1952, Green had ranked over one hundred tumours on the basis of their transplantability.

2.3 Inventing an Abstract Quantity

The key step I am interested in here is this invention of an abstract quantity to which transplantation provided access. Some, following Loeb, took this abstract quantity to be the "growth momentum":

> Clinically, the growth momentum of a tumor, i.e., the rate of enlargement, infiltration, and metastasis, characterizes the degree of malignancy of a neoplasm. It has been shown experimentally with animal tumors that growth momentum is likewise one of the most important factors governing transplantability, particularly heterologous transplantability. [. . .] Accordingly, the determination of heterologous transplantability of a tumor would provide a measure of its growth momentum and, hence, the degree of malignancy. (Towbin 1951, p. 716)

To make the analogy plain, "growth momentum" was the unobservable value to be measured (the equivalent of temperature), and the growth of the transplant was the signal (the equivalent of the height of the mercury column). For some, transplantability was a proxy to "growth momentum", for others it was a measurement of the "autonomy" of the tumours. In both cases, these abstract quantities were already loaded with both conceptual content and experience. Indeed, the notion of autonomy was already used to explain both developmental processes and carcinogenesis (see for instance the work of Hugo Ribbert or John George Adami, where both are explained as differential responses to "tissue tension"). Similarly, the notion of "growth momentum"[2] was central to Leo Loeb's biology (1945). Loeb himself was using transplantation as a measuring procedure, although his approach was more complex[3] Importantly, in both cases the abstract notion was used to explain a variety of phenomena, both natural and artificial, both normal and pathological.[4]

---

[2]The notion seems to come from demographics, where it became especially popular in the 1920s. Should this have been Loeb's inspiration, it would be yet another early example of what would become very recurrent analogies between cancer and socio-economics: Bolshevik cells, anarchist cells, etc. . . .

[3]Loeb considered his experiments to simultaneously measure different aspects of the phenomena which he called "differentials" (Loeb 1945). Although I believe that the present discussion could equally apply to his work, the presence of multiple quantities in the same reading makes the matter less straightforward.

[4]Because it took cancer and physiology as variations of the same causes, this approach was part of what Michel Morange called "the Regulatory Vision of cancer", in which "cancer was conceived as a disease of development." (Morange 1997, p. 6)

Given the diversity of transplantation procedures (host species, site of transplantation, assessment method, etc.), it should not come as a surprise that scientists produced different, conflicting classifications. This prompts the question of which transplantation system correctly tracks growth momentum (or the abstract quantity of choice). Here, we meet what has been the core problem of classical thermometry: without a direct access to temperature, how can one know which thermometer gives the right temperature?

3 Measurement and Theoretical Terms

But these questions already make an important assumption. Why did there have to be a single, true classification system – or, for that matter, a single, true temperature? Chang's "Ontological principle of single value", which states that "a real physical property can have no more than one definite value in a given situation" (Chang 2004, p. 90), simply displaces the question to why scientists believed that temperature was "a real physical property". At least part of the answer has to do with the fact that the notion of temperature already had an entrenched ancestry. It built upon both a long philosophical tradition that had already pre-conceptualized the notions of heat and "caloric", and an immediate, daily-life experience of variations in temperature.[5] Both of these loosely fitted the readings of thermometers, strongly suggesting that all were related to a common quantity.

To some extent, similar arguments can be made in the case of malignancy. In the quote from Towbin above, it is striking that the term "growth momentum" is associated to a disarraying variety of phenomena which, obviously, *could* point in different directions. But it was sufficient that they would align most of the time to postulate a common cause – after all, most scientific laws are *ceteris paribus*.

Moreover, the clinical context of xenotransplantation experiments already provided a very concrete notion of malignancy: the clinical outcomes of patients provided a grading of tumours. In a sense, therefore, the clinical could be understood as the (de facto) unobservable to which the instruments are giving access. In this context, transplantation experiments simply ought to approximate clinical outcomes. Indeed, scientists tinkered their transplantation procedures to approximate clinical knowledge.[6] Nevertheless, as will become obvious in the next section, scientists did not go all the way in this direction. Two important reasons can be

[5] As Chang writes: "human sensation serves as a prior standard for thermoscopes" (Chang 2004, p. 42)

[6] The test was used in some laboratories (Towbin 1951), but its sensitivity was criticized. Hence in the decades that followed, different methods were shown to make the hosts more receptive, including X-ray irradiation, cortisone treatment, and thymectomy, but the most important advance was certainly the discovery, in the early 1960s, of the *Nude* mouse mutant. Aside from its famous absence of hair, the nude mouse is characterized by its lack of a functional thymus and the corresponding massive reduction in T cells. As a consequence, it is largely unable to mount an

given for this. The first, supported by the passage from Greene quoted above, is that they were not just after a predictive system, but an explanatory, or at least exploratory one, enabling an investigation of the "mechanisms of autonomy" (Greene 1951, p. 902). In this context, approximating the clinical outcome has the value of highlighting departures from it, and therefore of stabilizing these departures as objects of explanation.

A second reason is the lack of repeatability of what we might call the "clinical measurements" of malignancy. Malignancy, understood as the ability for pathological, neoplastic growth, is a relational property of cells. Indeed, cells can to some extent become malignant just because of differences in the surrounding tissue (Bhowmick and Neilson 2004). This means that although each human tumour is associated to a medical history and clinical outcome, this history has many determinants that are external to the cancer cells. These can be due to the exact site and micro-environment of the tumour, the patients's constitution or genetic background, treatment history, etc., so that the correlation between the nature of the cancer cells and the medical outcome is messily statistical rather than deterministic. Nevertheless, experience strongly suggested that among the factors of malignancy were differences that were intrinsic to the cancer cells: hence the invention of abstract notions such as "growth momentum". In the absence of an independent mean to group tumours together (which transplantability was trying to offer), this meant that in practice each clinical outcome was yet another poor approximation of the tumour's "intrinsic malignancy". The invention of such an intrinsic property, because it makes it transportable (transplantable), enables the phenomena to simultaneously become a theoretical variable and an object of experimental study.[7]

To study cancer in a mouse, moreover in the highly artificial context of the laboratory, will never be the same as to study it in patients. However, the invention of an abstract property to which the measurement procedure would give imperfect access enabled the use of the mouse system to study the human system. In other words, the abstract concept provided a bridge between different material systems, by *assuming* that the material systems were simply two imperfect operationalizations of the same thing.

An analogous issue is ubiquitous in Chang's (2004) history of thermometry, although the question is never explicitly addressed: why did scientists need an abstract concept of temperature? Metallurgists were doing fine optimizing their furnaces with their solid thermometers, and scientists were designing steam engines on the basis of air thermometers. The only problem with purely operational concepts

immune rejection of the foreign tissues. In fact, even normal tissues successfully engrafted, but at that time Greene's idea of transplantability as a test of cancer was already forgotten.

[7]In fact, most of experimental biology is about *turning* context-sensitive features into capacities or stable properties. Arguably, the stunning success of the strategies of decomposition and localization in biology (Bechtel and Richardson 1993) is partly due to the fact that these strategies simultaneously provide understanding of the phenomena and constitute it as an epistemic thing (Rheinberger 1997).

is that they are unable to afford semantic expansion, and therefore knowledge gained, say, at the casual temperature range could not be easily transferred to very high or very low temperatures. The concept of temperature, abstracted from any operationalization, provided such a bridge.

3.1 Reaching a Thermometry Consensus

Hasok Chang's history of thermometry (Chang 2004) describes the many strategies with which scientists tried to stabilize the concept of temperature, many of which have an analog in the context of xenografts. For instance, a substantial part of the history of thermometry was aimed at establishing (or unsettling) "fixed points": the freezing point, melting point, boiling point, blood temperature, the first night frost, etc., up to the temperature of the cellars of Paris' Observatory (*id*., p. 10). The same could be said of xenograft experiments: Greene, Towbin, Loeb and others spent considerable efforts establishing fixed points. The most obvious is the inability of healthy differentiated tissue to grow, but more interestingly a variety of clinical characteristics (e.g. whether the tumour was metastatic) were believed, in clinical experience, to correlate with the abstract quantity. Chang's tale of how the fixity of fixed points was challenged, until fixed points had to be manufactured, would find many echoes here. But for the purpose of this paper, it is more useful to go to a later episode of Chang's history and briefly look at how the scientific community finally settled on what is nowadays taken for granted: absolute temperature.

In the end, most of the conundrum was solved when Thomson (Lord Kelvin) reasoned that the establishment of an absolute temperature required "a theoretical relation expressing temperature in terms of other general concepts", and relied on "the little-known theory of heat engines by the army engineer Sadi Carnot (1796–1832)" (*id*., p. 175).

> As Thomson was attempting to reduce temperature to a better established theoretical concept, the notion of mechanical effect (or, work) fitted the bill here. A theoretical relation between heat and mechanical effect is precisely what was provided by a theory of heat engines. (*id*., p. 175)

The first step was therefore to postulate an abstract temperature as defined by its theoretical (and quantitative) relationship with another abstract term, "work", which was linked to operational concepts through mechanics. Interestingly, the second step was then a "deliberate conflation" of this absolute temperature and of the temperature given by any thermometer (air, mercury, etc.): physicists assumed that the thermometers gave imperfect readings of this abstract quantity, and simply substituted one for the other in their formula (*id*., p. 214). Obviously, the fit was not perfect, but discrepancies allowed scientists to recalibrate their instruments, and engage in successive steps of approximation and recalibration which Chang characterized as "epistemic iteration": "point-by-point justification of each and every step is neither possible nor necessary, what matters is that each stage leads

on to the next one with some improvement." (*id.*, p. 215). While such iterations need not necessarily converge, when they do it vindicates both the instrument and the theoretical construction.

It is interesting to note that the theory provided both the motivation and the solution to the problem: scientists investigated thermometry to build a theory of temperature, and yet is it the theory which solved the problems of thermometry. There is no problem in this circularity: tools and theories that are rightly articulated gradually stabilize each other. But it means that there is considerable freedom in the starting point. Indeed, the theoretical relationship between work and temperature was enabling a quantitative theory of heat and an explanation of thermometry, therefore giving reasons to connect the different "operational temperatures". But at the same time, it provided ways of going from one thermometer to the other, therefore undermining the need for "the right thermometer".

3.2 Xenografts in the Cancer Stem Cell Framework

Nowadays, saying that a tissue grows because of its growth momentum is as explanatory as saying that opium makes one sleep because of its "vertu dormitive". The situation was certainly different for scientists thinking within the theoretical context of Loeb (1945). Hence the fact that the notion of the "growth momentum" of cancer tissues did not catch on is most surely related to the demise of the notion in developmental biology. In any case, both growth momentum or Greene's "autonomy" lacked tractable relationship with other notions, and this proved critical for the establishment of a theory-instrument articulation.

From the 1960s on, and especially until the 1990s, cancer research was unified around a different notion, dissociated from physiology: tumorigenicity. While the concept of tumorigenicity would deserve a history of its own, it is an umbrella term gathering so many heterogeneous meanings that its treatment would only distract from the present discussion. Instead, I would like to discuss a more recent episode of xenograft experiments. Because it bears a strong resemblance to the previous example, its analysis can benefit from the previous discussion.

At the turn of 2000s, strong analogy between physiology and pathology resurfaced in cancer research under the form of the Cancer Stem Cell (CSC) hypothesis. I shall only briefly summarize the CSC model here – for a more detailed discussion, see Blasimme et al. (2013), Visvader and Lindeman (2012), and Valent et al. (2012). Its core hypothesis is that cancer progression is driven by a small subpopulation of tumour cells with stem-cell-like properties. Like in normal tissues, only these cells are capable of infinite replication, and they therefore fuel a hierarchical tissue development. After the discovery that myeloid leukaemia followed such a model, a whole research programme developed with the aim of identifying and isolating such cells in other forms of cancer. The basic strategy is to divide the population in subpopulations according to some markers (typically on the cell's membrane,

so that cells can be sorted through antibody-based methods), and assess whether these subpopulations differed in terms of some measurement. Once more, mice were recruited as measuring devices, and once more, a variety of transplantation procedures resulted in conflicting measurements.

One of the best examples of this conflict is the controversy regarding melanoma stem cells. In the field of melanoma research, scientists have proposed to speak of melanoma-initiating cells (MIC) as an operational definition of CSC: MIC are cells which, when serially transplanted into an immuno-deficient mouse, are able to produce tumours recapitulating the heterogeneity of the original tumour. Strictly speaking, scientists are most often not measuring whether the cells are able to produce tumours, but to what extent, and therefore the injected cells are not said to be all CSC, but to be enriched in CSC.

A few years ago, Schatton et al. (2008) identified a sub-population of cells, ABCB5+ cells (cells expressing the ABCB5 antigen at their surface) enriched in what they claimed to be CSC. In order to test it against the operational definition of MIC, they transplanted ABCB5- and ABCB5+ populations of cells from a human tumour into NOD/SCID immunodeficient mice ("non-obese diabetic/severe combined immunodeficiency") and looked at the tumour progression. After 8 weeks, hardly any tumour grew in the first case, and the majority steadily grew in the second. In other words, only a small proportion of tumour cells, strongly enriched in the ABCB5+ population, were able to initiate and sustain new tumours. They published an enthusiastic letter to nature which was heavily cited, and for a time it was proclaimed that CSC had been identified in melanoma.

Some months later, Morrison's lab (Quintana et al. 2008) published a paper attacking these claims. The most important for the present discussion is that they tried the same experiments with an even more immunocompromised mouse (the NOD/SCID Il2rd-/- mouse) and obtained radically different results. Injecting single cells, they found that one out of four was able to initiate palpable tumours, and trying a wide range of markers, they were not able to correlate this with any signature. They therefore concluded that there was no proof yet that the CSC model obtained in melanoma, and that experiments seeking tumour-initiating cells should beware of relevant differences between the tumour environment in the patient and in the mouse host. The lesson, it seems, was that Schatton et al. (2008) had drawn a bad conclusion that was due to the particular mouse model they used, which happened to be unrepresentative of the human host. Quintana's paper was (and still is) a big success, being cited even more than the first, often as a methodological warning. Nevertheless, given how "unnatural" the dramatically immunocompromised mice are, there is still considerable debate as to which is the best (see for instance Civenni et al. 2011).

The mice, and in fact the whole experimental system, again acted as an instrument: they transformed an unobservable, yet causally relevant difference, into a visible signal, thus revealing this difference. But what difference exactly? Given the disagreements of two recipients, which signal faithfully informs us about tumorigenicity or "CSC-ness"?

The question becomes even more acute if we consider the rest of the story. Slightly more than a year after Quintana's paper, Schatton et al. (2010) published a follow-up paper in *Cancer Research*, apparently moving the topic: "Modulation of T-Cell Activation by Malignant Melanoma Initiating Cells". Taking the discrepancies between the two studies as a starting point, they addressed the question of why the difference between the mouse strains – the absence or presence of the interleukin-2 gamma receptor (Il2rg) – made such a difference to the apparent role of ABCB5+ cells. It turned out that ABCB5+ cells seem to block or reduce the proliferation of immune cells and the production of interleukin-2, thus modulating T-cell activity. Obviously, in a mouse which anyway lacks such an activity (and, as a matter of fact, that completely lacks interleukin-2 gamma receptors), one expects to find no difference between the subpopulations of cells. But in a mouse that has such an activity, only cells that are able to disrupt this mechanism can proliferate efficiently.

Assuming that ABCB5+ cells prove to also be more malignant in the case in humans, one might argue that the first model (the least immunocompromised) was a better model. However, Morrison's group would probably point out that this malignancy is not due to the tumorigenicity of the cells per se. But *on what ground can one exclude phenomena as part or not of such an abstract property*? On closer inspection, which instrument is the best depends on what it is that we wish to measure – in this context, on the understanding one has of tumorigenicity or CSC-ness.

Tumorigenicity is the capacity to form tumours and sustain growth, but in the presence or in the absence of an immune pressure? One the one hand, human tumours do not develop "in the void": cancer patients are seldom so immunodeficient, and immune response is an important part of cancer development and of variability in outcomes. A notion of tumorigenicity independent of this pressure seems to be an idealization that lacks practical relevance. On the other hand, it seems scientifically worthwhile to isolate the different components influencing the malignancy of cancer cells, so that we might want to exclude the effects of the immune system: tumorigenicity is one thing, evasion of immune surveillance is another. The problem with this reasoning is that many other causally relevant elements (many ways through which some cells might be more tumorigenic than others) could also be excluded. Therefore, one can legitimately ask why excluding this and not other causally relevant elements. The only reasoned answer one can provide has to be linked to the adoption of a theoretical framework[8] – or what one could call the "theoretical grounding" of operations or instruments.

[8] See also Griesemer (1992) for discussion of how the appropriateness of a tool is necessarily linked to the adoption of a theoretical framework.

3.3 CSC and Theoretical Grounding of Operations

In the first paper that I briefly described in the previous section, Quintana et al. (2008) reduce the CSC model to tumorigenicity: "the cancer stem-cell model has suggested that only small subpopulations of cancer cells have tumorigenic potential" (Quintana et al. 2008, p. 593). There has been a general tendency, especially in the field of melanoma, to avoid the abstract talk of CSC in favor of the operational talk of melanoma-initiating cells – or cells that initiate melanoma when transplanted into an immuno-deficient mouse. Likewise, participants of the 2011 Working Conference on CSC have explicitly tried to split the conceptual and operational meanings (Valent et al. 2012) in order to avoid a conflation of the two. However, severing the connection between the two is equally problematic. An exclusive focus on tumour-initiating potential would be like a focus on the height of the mercury column: while it might be useful locally, it does not allow semantic extension. The abstract concept does. The CSC framework can potentially mediate between the material contexts. But this means that the problem of selecting the "right" xenograft model can only be solved if one has at least a tentative theoretical understanding of what the instrument should measure.

I believe the CSC framework can succeed where the notion of growth momentum has failed precisely because its meaning has theoretical implications which are not reducible to the operational definition of CSC. The seminal findings of Bonnet and Dick (1997) was not that some leukemic cells were more tumorigenic than others, but precisely that those were the cells possessing stem-cell like characteristics ("the differentiative and proliferative capacities and the potential for self-renewal" Bonnet and Dick 1997, p. 730). In doing so, it established a parallel between cancer and normal development, suggesting that the physiological differentiation hierarchy can shed light on the dynamics of cancer. In other words, it also poses additional constraints as to the kind of measurements that ought to be linked to it. Schatton's findings of the modulation of T-cell activity by cancer cells may be extremely relevant for an understanding of cancer, but it has no physiological counterpart and is unrelated to the tissue hierarchy. As such, it is irrelevant to the identification of CSC. It is not biologically or clinically irrelevant, but irrelevant to what it is that the xenograft was supposed to measure. Insofar as the xenografts are used for the identification of CSC, the question of the "right" assay can only be answered with respect to the theoretical meaning of CSC: beyond its operationalizations and with full attention to the theoretical relationships it entertains.

4 Conclusion

There are several differences between thermometry and xenografting, or between thermodynamics and the CSC framework. The fact that physics is quantitative is perhaps the most important one: ordinal measurements of growth momentum

were not quantitative in the sense that they allowed no meaningful arithmetic operation between measurement results. Nevertheless, as I have tried to show, there are important similarities in the epistemological problems encountered, and the comparison can yield insights into the current problem of the choice of xenograft host. But I would now like to conclude with more general observations.

The comparison reveals that Chang's notion of "epistemic iteration" is not limited to quantitative cases. Instruments are often represented as sorting devices from which emerge classifications of reality (see for instance Buchwald 1992). At the same time, entertaining the full multiplicity of instruments (or procedures), and hence of competing classifications, would be counter-productive. What I have tried to highlight is that theoretical frameworks, however preliminary or vague they are, are needed to restrict this plurality. The first step is to *assume* an identity between an operational concept and a theoretical concept, and the importance of such bootstrapping assumptions was also emphasized in other fields such as experimental psychology (Sullivan 2008; Feest 2010). From that point on, the process is one of gradual correction of both terms to resolve inconsistencies. In the case of xenograft experiments, the use of increasingly immunodeficient mice, or of humanized mice (Maugeri and Blasimme 2011), are gradual corrections of this kind. What I have been trying to emphasize, however, is that a lack of convergence is not a failure of the instrument: it is a failure of the whole articulation between instrument and theory.

Until a theoretical framework has been shown to be satisfactory, there is neither an epistemic ground nor even a reasonable motivation for operational monism. And once the theoretical framework is complete, such as in the case of thermodynamics, there is no more the need for operational monism – for the "true" temperature, since the values of one thermometer can be converted to those of another. It is in between these two moments that theoretical terms are the most productive. It is precisely because they are operationally vague, but not as vague as to defy transposition (we could say that they are operationally *suggestive*) that they allow mediation between material contexts. Yet to avoid a trivializing flexibility, their meaning has to be restricted through relations to other theoretical terms.

References

Bechtel, W., and R.C. Richardson. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge: MIT Press.

Bhowmick, N., and E. Neilson. 2004. Stromal fibroblasts in cancer initiation and progression. *Nature* 432: 332–337.

Blasimme, A., P. Maugeri, and P.-L. Germain. 2013. What mechanisms can't do: Explanatory frameworks and the function of the p53 gene in molecular oncology. *Studies in History and Philosophy of Biological and Biomedical Sciences* 44(3): 374–384. doi:10.1016/j.shpsc.2013.02.001.

Bonnet, D., and J. Dick. 1997. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Medicine* 3(7): 730–738.

Buchwald, J. 1992. Kinds and the wave theory of light. *Studies in History and Philosophy of Science Part A* 23: 39–74.

Chang, H. 2004. *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Civenni, G., A. Walter, N. Kobert, D. Mihic-Probst, M. Zipser, B. Belloni, B. Seifert, H. Moch, R. Dummer, M. van den Broek, and L. Sommer. 2011, April. Human CD271-positive melanoma stem cells associated with metastasis establish tumor heterogeneity and long-term growth. *Cancer Research* 71(8): 3098–3109.

DeVita, V.T., and E. Chu. 2008, November. A history of cancer chemotherapy. *Cancer Research* 68(21): 8643–8653.

Ewing, J. 1919. *Neoplastic diseases, a text-book on tumors*. Philadelphia: W.B. Saunders.

Feest, U. (2010). Concepts as tools in the experimental generation of knowledge in cognitive neuropsychology. *Spontaneous Generations: A Journal for the History and Philosophy of Science* 4(1): 173–190.

Funk, C. 1915. The transplantation of tumors to foreign species. *The Journal of Experimental Medicine* 21(6): 571–3.

Germain, P.-L. forthcoming. From replica to instruments: Animal models in contemporary biomedical research. *History and Philosophy of the Life Sciences*.

Greene, H.S. 1951. A conception of tumor autonomy based on transplantation studies: A review. *Cancer Research* 11(12): 899–903.

Greene, H.S. 1952. The significance of the heterologous transplantability of human cancer. *Cancer* 5(1): 24–44.

Greene, H.S.N. 1948. Identification of malignant tissues. *JAMA: The Journal of the American Medical Association* 137(16): 1364–1366.

Griesemer, J.R. 1992. The role of instruments in the generative analysis of science. In *The right tools for the job: At work in the twentieth century life sciences*, 47–76. Princeton: Princeton University Press.

Hanau, A. 1889. Erfolgreiche experimentelle Übertragung von Carcinom. *Fortschritte der Medizin* 7: 321–339.

Hekzog, M. 1902. On tumor transplantation and inoculation. *The Journal of Medical Research* 1(6): 74–84.

ILAR and NRC. 1998. Biomedical models and resources: Current needs and future opportunities. Technical report, Committee on New and Emerging Models in Biomedical and Behavioral Research, Institute for Laboratory Animal Research, Commission on Life Sciences, National Research Council.

Loeb, L. 1945. *The biological basis of individuality*. Springfield/Baltimore: Charles C. Thomas.

Maugeri, P., and A. Blasimme. 2011. Humanised models of cancer in molecular medicine: The experimental control of disanalogy. *History and Philosophy of the Life Sciences* 33: 603–622.

Mayet, M. 1902. Production du Cancer chez les Rats blancs par Introduction dans leurs economies des Substances constituantes des Tumeurs malignes de l'Homme. *Gazette hebdomadaire de médecine et de chirurgie* 6: 64–68.

Morange, M. 1997. From the regulatory vision of cancer to the oncogene paradigm, 1975–1985. *Journal of the History of Biology* 30: 1–29.

Murgia, C., J. Pritchard, S. Kim, A. Fassati, and R. Weiss. 2006. Clonal origin and evolution of a transmissible cancer. *Cell* 126: 477–487.

Novinski, M. 1876. Zur Frage über die Impfung der krebsigen Geschwulste. *Zentralbl. Med. Wissensch.* 14: 790–791.

Quintana, E., M. Shackleton, M.S. Sabel, D.R. Fullen, T.M. Johnson, and S.J. Morrison. 2008. Efficient tumour formation by single human melanoma cells. *Nature* 456(7222): 593–598.

Rheinberger, H.-J. 1997. *Toward a history of epistemic things: Synthesizing proteins in the test tube. Writing science; variation: Writing science*. Stanford: Stanford University Press.

Schatton, T., G.F. Murphy, N.Y. Frank, K. Yamaura, A.M. Waaga-Gasser, M. Gasser, Q. Zhan, S. Jordan, L.M. Duncan, C. Weishaupt, R.C. Fuhlbrigge, T.S. Kupper, M.H. Sayegh, and M.H. Frank. 2008. Identification of cells initiating human melanomas. *Nature* 451(7176): 345–349.

Schatton, T., U. Schütte, N.Y. Frank, Q. Zhan, A. Hoerning, S.C. Robles, J. Zhou, F.S. Hodi, G.C. Spagnoli, G.F. Murphy, and M.H. Frank. 2010. Modulation of T-cell activation by malignant melanoma initiating cells. *Cancer Research* 70(2): 697–708.

Snell, G.D. 1964. The terminology of tissue transplantation. *Transplantation* 2: 655.

Sullivan, J.A. 2008. The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese* 167(3): 511–539.

Towbin, A. 1951. The heterologous transplantation of human tumors. *Cancer Research* 11: 716–722.

Valent, P., C. Eaves, D. Bonnet, R. De Maria, T. Lapidot, M. Copland, J.V. Melo, C. Chomienne, F. Ishikawa, J.J. Schuringa, G. Stassi, B. Huntly, H. Herrmann, J. Soulier, A. Roesch, G.J. Schuurhuis, S. Wöhrer, M. Arock, J. Zuber, S. Cerny-Reiterer, H.E. Johnsen, and M. Andreeff. 2012, October. Cancer stem cell definitions and terminology: The devil is in the details. *Nature Reviews Cancer* 12(11): 767–775.

van Fraassen, B.C. 2008. *Scientific representation: Paradoxes of perspective*. Oxford: Clarendon Press.

Visvader, J., and G. Lindeman. 2012, June. Cancer stem cells: Current status and evolving complexities. *Cell Stem Cell* 10(6): 717–728.

Witkowski, J.A. 1983. Experimental pathology and the origins of tissue culture: Leo Loeb's contribution. *Medical History* 27(03): 269–288.

Zondek, B. 1928. Die Schwangerschaftsdiagnose aus dem Harn durch Nachweis des Hypophysen-vorderlappenhormons. *Die Naturwissenschaften* 51: 1088–1090.

Developmental Explanation

Veli-Pekka Parkkinen

1 Introduction

An explanation-seeking question "why does a system S exhibit a behavior B" is ambiguous, which means that it can be answered in many ways. One can give an etiological account that describes how the system's interactions with its environment triggered the behavior. Or one can describe the components that the system is made of, to explain how the behavior is realized in the system. The latter is an explanation by appeal to constitution. Whence etiological explanations describe what triggered a behavior of a system, constitutive explanation tells how the system is capable of exhibiting the behavior in question. Furthermore, when offered a constitutive explanation of a capacity of a system in terms of its components and their relations, one can ask for further information about how the system was built or developed to have a certain constitution at a certain point in time. This paper analyzes explanations of the last-mentioned type, where the explanandum is provided by the changing constitution of a developmental system.

I will investigate the structure of such developmental explanations by using a contrastive-counterfactual framework. This account takes explanations to be tracking change-relating dependencies. Thus, explanation involves considering hypothetical scenarios where some factors vary while others remain stable. Those variables X that if changed in some background-conditions, would make a difference with respect to some other variables Y, provide the putative explanans for the variables Y. To ground explanations, these dependencies should support manipulations of the explanandum variables. In this view, explanations always address a contrast: to provide an explanation is to show why it is the case that *y*,

V.-P. Parkkinen (✉)
Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Blindernveien 31, 0313 Oslo, Norway
e-mail: v.p.parkkinen@ifikk.uio.no

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_11,

rather than y'. The explanans should point out factors that would need to be changed in order to actualize the contrasts specified in the explanandum. I use this analysis to identify similarities and differences in etiological, constitutive and developmental explanations.

Explaining development involves showing how characteristics of later developmental stages are produced as the manifestation of the system's earlier developmental capacities. This presupposes an account of how these capacities depend on the properties of the system's components at each stage. Explaining development therefore involves addressing two types of explananda: etiological explanandum about the manifestation of capacities, and constitutive explanandum about the realization of those capacities. However, no explanation is complete in the sense that it would describe every interdependent aspect of the developmental process at once: actual explanations typically focus on a specific aspect of development.

The analysis of developmental explanation can be employed to make sense of the problem of reductionism in developmental explanation. Alexander Rosenberg (1997, 2006) has famously argued that developmental biology only became explanatory when the tools for uncovering the underlying molecular processes became available, and that tracking the changes in the molecular composition of an organism will eventually provide answers to all the questions there are about development. This controversial claim has since been contested, but the specific target and scope of the arguments is unclear, due to the fact that the content of the notion of explanation is left to reader's intuition.

I employ the contrastive-counterfactual framework to better understand in what sense developmental biology is, and in what sense it is not, a reductionistic project. Explicating the structure of developmental explanations removes ambiguity about what kind of information is requested for when explaining aspects of developmental phenomena. An actual explanation typically addresses one of the possible explanatory tasks at a time. Some developmental explanations, such as those in developmental psychology, focus on demonstrating how capacities of a mature organism depend on the exercise of similar but limited capacities earlier in development, without explicit reference to the underlying constitution of the system. In comparison, developmental biology often focuses on explaining a specific developmental capacity constitutively, such as in the canonized textbook examples like *Drosophila* axis formation or the gastrulation of the sea urchin embryo, where the capacity of an organism to undergo a specific structural change is shown to be due to the properties and organization of relevant constituents of the system at a specific point in development.

This might give the impression that the true explanatory power in developmental biology lies exhaustively in describing the particular configuration of the constituents of an organism at each stage. But the framework of developmental explanation permits also asking questions about the origins of the components and their organization. Answering these questions involves tracking factors that might not be included in a description of changes of the actual molecular constitution of the system over time. Laubichler and Wagner (2001) draw attention to questions of this

type in their critique of Rosenberg's reductionist position, and discuss specific cases that ought to demonstrate that Rosenberg's molecular reductionism is incapable of handling certain legitimate explananda of developmental biology. I discuss two types of cases presented by Wagner and Laubichler, namely the many-one and one-many relations between molecular components and developmental outcomes, and evaluate whether these cases are able to establish that there is more to developmental biology than the description of molecular properties and interactions over time.

The structure of the paper is the following. First, I explain the basic ideas of a contrastive-counterfactual theory of explanation and interventionism, and describe how etiological and constitutive explanation can be understood in these terms. Then I describe how explaining development fuses these two aspects, and characterize the framework of developmental explanation. Finally, I evaluate whether molecular reductionism can address the explananda that can arise in a developmental framework, as exemplified in Laubichler and Wagner's counterexamples to Rosenberg's view. The concluding chapter summarizes the main points.

2 Interventionist Theory of Explanation

Explanatory controversies often suffer from an ambiguity concerning the target of the explanation. A way to clarify this is to think about explanations as answers to contrastive questions, so as to make explicit the explanandum. In this view a request for an explanation is of the form "Why P rather than Q?", and an explanans should provide information that allows one to formulate an answer "P rather than Q because F rather than G" (Achinstein 1983; Garfinkel 1981; Schaffer 2005). This suggests that the information needed for an explanation is about dependence-relations.

Not all dependence-relations support explanations. General criteria for what is needed for a dependence-relation to be explanatory can be given with the help of the idea of manipulability, and the technical notion of an intervention. This is the starting point of the now widely adopted and discussed interventionist theory of explanation due to James Woodward (2003). Woodward proposes his theory as an account of specifically causal explanation, but here I use the resources of the theory to investigate differences between different kinds of change-relating dependencies in terms of what kind of manipulations they support, and what kind of contrastive questions one can address given knowledge of these dependencies.

According to Woodward, "we are in a position to explain when we have information that is relevant to manipulating, controlling, or changing nature, in an 'in principle' manner of manipulation .... We have at least the beginnings of an explanation when we have identified factors or conditions such that manipulations or changes in those factors or conditions will produce changes in the outcome being explained." (Woodward 2003, pp. 9–10)

More specifically, explanation consists in describing function-like dependencies between relata that can be represented as values of variables. Explanation should provide counterfactual information about the behavior of the explanandum variable

by showing how hypothetical changes in the value of the explanans would relate to changes in the value of the explanandum. Explanations thus create understanding by answering *what-if-things-had-been-different?* questions about the explanandum (Woodward 2003; Ylikoski and Kuorikoski 2010).

The hypothetical changes that Woodward considers as providing the counterfactual content of explanation are *interventions*. Intervention is thought to be an exogenous process that directly sets the value of the (or value of some of the) explanans variable(s), while not directly affecting the explanandum variable in any way. Any change in the explanandum variable that is associated with an intervention must come via changing the explanans variable, i.e., the intervention must not directly cause the change in the explanandum variable, or influence it via some other route than through changing the explanans variables (Woodward 2003, pp. 98–99). Explanations should be such that they can be used to predict what would happen to the explanandum if interventions were to change the explanans. This relation ought to be invariant to a certain degree, meaning that the functional dependency between explanans and explanandum should hold for at least some values of the explanans variables set by interventions. Since invariance admits degrees, explanations need not mention any completely universal generalizations such as laws (Woodward 2003, pp. 249–251).

Explanations are contrastive in both the explanans and explanandum, claiming that "$Y = y_1$ rather than $y_2, \ldots, y_n$, because $X = x_1$ rather than $x_2, \ldots, x_n$", where possible values of X and Y form the relevant contrast classes for the explanans and the explanandum. Contrastivity coupled with the idea of interventions provide the criteria for explanatory relevance. When asking for an explanation for the occurrence of y_1, we do not ask to know why y_1 occurred *simpliciter*. Rather we want to know why y_1 occurred instead of some other conceivable outcome. The state of affairs that y_1 is then compared to actual or imagined situations where some other states of affairs $y_2, \ldots, y_n$ hold. The task of an explanation is to pick out the differences in these situations that correspond to variation in the explanandum. Actual explanations are thus selective, that is, they should mention only the factors that would vary corresponding to the intended contrast values of the explanandum (Achinstein 1983; Garfinkel 1981; Woodward 2003).

These general features of explanatory relevance can now be used to characterize causal and constitutive explanation, and developmental explanation as a combination of the two.

3 Etiological Explanation

One way to read an explanation-seeking question "Why does S exhibit behavior B?" is to rephrase it as "What triggered S to B (at a time t)?".

This requests for contrastive information of specific kind. Here S that exhibits B is compared to a real or imagined similar system S′ that did not engage in B-ing at time t, and the explanation consists of describing the factors that vary corresponding

to this difference in the behavior of S compared to S′. Since S and the contrast system S′ are taken to be similar, the variation that explains the difference in B-ing must lie outside the system itself.

The explanatory factors should be more than mere correlates to the system's behavior in its environment; they should be such that when changed by interventions, they would make a difference to whether S will exhibit B or not. An answer therefore would state that S exhibits B rather than B′ because of some conditions or factors C rather than C′, that support manipulations between B-ing and B′-ing. Note that the explanandum here is the behavior B, so it is best understood as an event, or a process, that takes time. The factors that explain the event or process should be of a comparable metaphysical kind, i.e., they should be events that trigger S′s B-ing or start up the process. Thus it is natural to think of the dependence between the explanans and the explanandum as causal; earlier events explain later ones, or start-up conditions of a process explain its outcomes (Salmon 1984; Craver 2007, pp. 73–75, Ylikoski 2013). Etiological explanations have such form: The causes of S exhibiting B are located in the causal history of S (Salmon 1984; Craver 2007, pp. 73–75). A typical explanandum is some behavior of a system that is made understandable by referring to exposure to environmental factors that triggered that behavior.

To detect etiological explanantia, one intervenes to change a putative cause, and detects a change in the behavior of the system. In order to reveal explanatory relations, the testing intervention must not directly change the explanandum, i.e., the intervention must be such a process that it changes a putative cause of S′s B-ing, but does not directly change factors that make up S′s B-ing (Craver 2007, pp. 96–98, Woodward 2003, pp. 98–99).

Underlying an etiological question is a question of how the is system capable of exhibiting the explanandum behavior. For instance, let's say that the behavior of migrating birds is explained by the change in the period of daylight. An underlying question asks how the birds are able to detect changes in daylight, how they navigate on their journey, etc. The simple etiological explanation presupposes that there is an answer to underlying questions of this kind. To illuminate an etiological explanation, we can ask about the features of the system that enable it to interact with its environment in the required way and give rise to the system's ability to exhibit the behavior given suitable triggering conditions. Answering these questions creates additional understanding of the explanandum phenomenon by showing how the effects of the etiological factors on the system behavior could be prevented or modified by modifying parts of the system itself.

4 Constitutive Explanation

The question "Why does S exhibit B" can also be read as "How does S realize B-ing?"

This is a question about the constitution of the system.

The target of an explanation that addresses this question is a capacity of a system. The explanandum is a description of what the system would do in such-and-such a causal environment (Cummins 2000; Ylikoski 2013). The contrast in the explanandum, then, is between causal possibilities: a behavior B that a system S would exhibit in some environment is contrasted to behavior B′ that a real or imagined comparison system S′ would exhibit *in the same* causal environment. The task of an explanation is to pick out the differences in such situations that correspond to the difference in whether a system would exhibit B or B′. Given that the conditions external to the system are considered to be stable, the source of variation must be within the system. Those factors within the system that, if varied, would change what the system would do in a given causal environment are the ones that belong to the explanans of a constitutive explanation.

In giving etiological explanations, we are trying to understand manifestations of the capacities of a system by tracking changes in the system's environment. By contrast, constitutive explanation seeks to understand what in the system makes it possible for it to respond to certain causal environments in certain ways. To detect constitutive explanatory relevance, we can look for component behaviors that coincide with the manifestation of the system's capacities. These should be more than mere correlates of system behavior in the behavior of components. Observing that the behavior of a component undergoes a synchronous change when a capacity is manifested is a way to detect candidate constitutive explanantia, just as observing correlations between events in the environment and system-behavior is indicative of possible etiological explanantia, but not sufficient for explanation. Component properties that figure in the constitutive explanation should support manipulations of the capacity in question. That is, manipulating the constituents of a capacity to B should make a difference with respect to how the system would respond to the causal conditions that normally trigger B (Craver 2007, pp. 139–141, pp. 145–147).

To detect constitutive relevance it is required to intervene to change a component while exposing the system to conditions that usually trigger the capacity of interest, and detect differences in how the capacity is displayed (Craver 2007, pp. 145–147). The component properties and their organization that could be harnessed to manipulate the display of a capacity in a fixed causal environment are the ones that figure in a constitutive explanation. Analogous to the requirement that an intervention on a putative etiological cause must not directly change the effect, an intervention aimed at revealing constitutive dependence must not be such that it involves varying any etiological causes that would directly change how the system behaves. The environment must thus be held fixed.

For the purposes of the following discussion, certain parallels and differences between etiological and constitutive explanation should be mentioned. Both explanations track relations of dependence that support manipulations. Both types of explanations provide understanding by facilitating contrastive inferences about the explanandum. Both types of explanations involve information about what would happen if some factors were to vary while others remain stable (Ylikoski 2013). In etiological explanation, the explanatory request is about the conditions that trigger the manifestation of capacities of a system: we have to consider a system that

exhibits B in contrast to a system with a similar structure that does not exhibit B. Constitutive explanation has us imagining what would happen to the system's abilities to exhibit B if some components of the system were changed. With respect to explaining an instance of B-ing, the constitutive explanation contrasts a system S that exhibits B in some circumstances to a different system S′ that does not exhibit B in the same external circumstances, and identifies the factors that differ in the system's constitution as the explanans.

Constitutive explanations differ from etiological explanations in the metaphysical characteristics of the relation supporting the explanation. Etiological explanations trade on causation, which is typically associated with events or time-consuming processes. Explanations are supposed to track this asymmetry of causes and effects in time, and therefore explain distinct events with antecedent cause-events, or outcomes of processes by their initial conditions. By contrast, constitution is a relation between a system and its parts, which means that the relata are not distinct entities nor separated in time (Ylikoski 2013).

These metaphysical differences amount to noteworthy differences in explanatory relations. The relata of a causal dependence afford independent interventions – we could intervene on a cause (effect) variable while holding fixed the effect (cause) variable with another intervention – but things linked by constitution do not. Given that it is conceivable that many different configurations of components might give rise to the same capacity, some properties of components may be manipulated in some ways without necessarily changing a capacity of a system. But manipulating a capacity of a system is the same thing as changing some of its components, therefore any change in a capacity must involve some change in components. This asymmetric dependence can be captured in terms of supervenience of the system-properties from its parts, though it should be noted that supervenience in itself is clearly not enough for explanation.

Finally, we should notice that according to the characterization of explanation endorsed here, causal explanations are not confined to any level of explanation, nor do they automatically follow any particular direction in a hierarchy of levels, no matter how this is conceived. A putative causal relation can be investigated between any factors that can be represented as causal variables in the sense that an intervention can be defined with respect to them; the assignment of variables to different levels doesn't matter. Whether the required manipulability relation between the variables actually holds either way is an empirical question. When it comes to constitution, matters are different. Given supervenience between constituents and a constituted whole, any change in the properties of a system necessarily involves a change in its constituents. This means that for variables that represent properties of a system and properties of its parts, it is impossible to define an intervention on a system-variable with respect to the variable that represents the properties of the parts. Changing a system is directly changing its constituents. Constitutive explanation thus flows bottom-up by fiat, when one considers a level-hierarchy based on part-whole relations. This also rules out some claims about downward causation, i.e., the idea of the determination of component-behavior by the same system that hosts the component. An explanatory setting where a system explains its parts in this way

is conceptually confused. There's nothing wrong in saying that being a part of a system has an effect on the behavior of an entity, but that effect is a result of being in a causal environment formed by the other parts of the system, and should be studied as such.

5 Developmental Explanation

Given the view on explanation endorsed here, explanation is an investigation into how hypothetical changes in the explanans would manifest as changes in the explanandum, given some background-conditions that are taken to remain stable. What is considered to vary and what isn't depends on the intended contrast in the explanation-seeking question. According to this analysis, etiological and constitutive explanation differ with respect to the locus of variation in the counterfactual situations that we are considering. Etiological explanation traces variation in system-behavior to variations in the system's external conditions when the constitution of the system and its causal capacities are considered fixed. In comparison, constitutive explanations explain those capacities, and traces hypothetical variation in the capacities to variation in the system's components when the environment is considered fixed.

In addition to asking for information about etiology or constitution, one can ask a third type of question with respect to a system exhibiting some behavior B: How did a system *acquire* the capacity to B, i.e. how was it built or developed so that it has a constitution that endows it with the capacity to B?

In a developmental process, there is no external agent responsible for the assembly of the system. The explanation thus ought to account for how each consecutive developmental stage is produced as a result of the system's developmental capacities in its earlier stages. This is a causal process: change the system's developmental capacities or their manifestation at one point in time, and the system will have different characteristics in a later stage. But in contrast to etiological explanation, in developmental explanation the explanantia are partly located within the system itself.

One way to describe development is to identify a series of changes in the causal capacities of a system over time (Ylikoski 2013). As causal capacities constitutively depend on the system's components and their organization, changes in the components must underlie the changes in the capacities. The explanation ought to show how the explanandum is a product of a *causal* process of changes in the system's constitution, whereby the constitution of the system at each stage endows it with the required capacities to proceed to the next stage.

Since explaining development combines etiological and constitutive explanation, it requires considering effects of variation in both the environment and the components of the system. However, no explanation can at once make understandable every aspect of a developing system. Instead, a meaningful request for explanation concerns only some aspects of this process at a certain point in developmental time.

Explaining developmental phenomena involves active choices of isolating one aspect of the total process as the explanandum. This is nothing special, nor does it render development beyond scientific understanding. Complex phenomena always need to be addressed from many different vantage points that each make their own simplifications (Wimsatt 2007). It does, however, invite the error of reifying a specific choice of perspective as the one and only privileged point of view. This will be discussed in the next chapter.

Ylikoski (2013) characterizes developmental explanations as involving causal capacities both as explananda and explanantia. This seems to be true of some cases, e.g., when language acquisition is modeled as a process of scaffolding the full mastery of a language through the exercise of more limited language-abilities in earlier stages of development. This would explain the end state of development by showing how the acquisition of the mature language-capacities depend on the right kind of environmental input and behavior of the system in the earlier stages. However, the same is not true in many typical examples of contemporary developmental biology explanations, which is what I want to consider here.

In actual explanations of developmental biology, the focus is usually to account for some changes in the constitution of the system, and not directly in the new capacities that the organism acquires through such changes. A textbook example for a developmental biology explanation is the formation of a segmented body plan in the *Drosophila* embryo. Here the primary objective is to understand how the segment-boundaries are created and what in the undifferentiated zygote determines its differentiated future geometry, not the new capacities that the segmented body plan supports.

As a combination of etiological and causal explanation, developmental explanation might give the impression of an ordinary mechanistic explanation: it describes in detail a causal process by describing the components involved in it (Machamer et al. 2000; Glennan 2002; Craver 2007). But this is not quite right. Mechanistic explanations explain the behavior or output of a system in terms of the operation of an underlying stable causal structure. By contrast, development involves changes in the very causal structure that is invoked as the explanans, after which it seizes to constitute the "mechanism for" that developmental step (McManus 2012).

The task of the explanation is to describe how it is possible for the organism to undergo this change in its components. Biologists want to understand what in the organism's earlier stages made it possible for the novel structures to be formed. This amounts to giving an account of the relevant constituents that endow the organism with the capacity to produce the developmental change of interest. The main target of the explanation, then, is often a capacity or a bundle of capacities that the organism possesses at a certain point in development. Understanding a specific developmental capacity always involves giving information of the constitutive kind. Adding a developmental perspective to such an explanation means to think of the components themselves as a causal product of the developing system's earlier capacities. Taking a developmental perspective, the contrast in the explanans of a constitutive explanation can be taken as an explanandum in itself.

To summarize, this is what I take it to mean to look at properties of a system from a developmental perspective: an explanatory setting for investigating the capacities of a system, where the existence of stable configuration of components is not taken for granted as a pre-existing fact, but can be taken up as an explanandum of its own.

6 Reductionism and the Desiderata for Developmental Explanations

Explaining biological development involves asking questions about how certain capacities of an organism are realized, and how in turn these realizers are made and configured in the developing system itself. Answering these kinds of questions is the desideratum of developmental biology.

Alex Rosenberg has famously argued that explaining biological development must involve an appeal to underlying molecular causes, and a molecular account suffices to explain everything about development (Rosenberg 1997, 2006). Rosenberg's claims have since been contested (e.g. Frost-Arnold 2004; Keller 1999; Laubichler and Wagner 2001; Soto et al. 2008) but the real purchase of the arguments on either side is somewhat unclear, because the notion of explanation remains unspecified. This situation can only be attenuated by giving an explicit characterization of the structure of developmental explanation. I will now evaluate both of Rosenberg's claims, as well as a critique put forth by Laubichler and Wagner (2001), that hinges on what the explanatory framework of developmental biology is taken to be.

Before proceeding, we should note that explanation as understood here does not require a derivation of the explanandum from the explanans. To explain something is to demonstrate a change-relating dependence that supports manipulations. Therefore, to explain system-level outcomes from component properties does not require that generalizations describing system-behavior be derivable from generalizations describing component behavior, nor that the vocabulary that is used to describe system-behavior be translatable into a vocabulary that describes the components. Thus, I will refrain from discussing reduction in the sense that has to do with deducibility and translatability between theories or vocabularies, since these questions don't matter to the question of whether a molecular account explains something or not. The schism between the participants of the reductionism debate is about whether or not a description of the particular molecular components that realize a developing system always suffices for explanation – this turns into a question of whether said details would always be the primary difference-makers for various phenomena one can take as the explananda in the context of development.

Rosenberg claims that developmental biology became truly explanatory only when a molecular characterization of the components that underlie biological functions became available. Before this, putative explanations of developmental biology were pseudo-explanations that merely reassert the phenomenon to be understood by attributing developmental dispositions to parts of an organism, akin to the way of "explaining" the sedative effect of opium by its "dormitive virtue" (Rosenberg 2006, p. 58).

Rosenberg illustrates this explanatory vacuity of early developmental biology with the concept of an organizer, which Spemann and Mangold employed to make sense of certain transplantation experiments. Grafting this region of a developing newt embryo to a random site in another embryo induces the development of a second embryo from that site. This region was then dubbed the organizer, due to its manifest power to determine the site where the development starts (Rosenberg 2006, pp. 58–59). However, knowledge of the organizer's power to induce development is hardly a good *explanation* of development, Rosenberg argues. To really understand what the organizer does, one needs other type of information than merely the knowledge that organizers induce development given a certain cellular environment.

To assess the explanatory merits of the organizer concept, we can ask if knowing the causal powers of the organizer puts us in a position to (hypothetically) manipulate aspects of the developing embryo. Using knowledge of the organizer, it is possible to design manipulations that result in the development of a partial embryo in an environment where it would not normally appear. One can therefore explain, by reference to the position of the organizer, why this part rather than that part of a homogeneous newt embryo began to develop into a differentiated embryo. However, this leaves us in the dark with respect to what should be manipulated in order to change any specific aspects of the embryos that develop from those sites. The organizer fails to point out potential targets for intervention with respect to the details of developmental processes.

The organizer-concept points to a developmental capacity – the capacity to create differentiated structures according to certain geometry – and identifies a region of the embryo that has a special role in how this capacity is achieved. Further understanding of the developmental events related to the organizer requires more detail about how this capacity is realized. We would need to know what exactly would have to be changed in order to change this capacity. This explanation is a constitutive one, the difference-makers for a capacity are some components that the system is made of. Although the behavior of the organizer region reveals interesting phenomena related to development, the concept has fairly limited explanatory power with respect to the desiderata of developmental biology. A real request for explanation here asks for a constitutive explanation of the capacity that is being referred to, and such an account is a bottom-up explanation by appeal to the constituents of the system.

But these kinds of constitution-questions are not the only ones that developmental biology asks. Laubichler and Wagner (2001) accuse Rosenberg of confusing the explanatory agenda of developmental biology, and cite specific cases that show why Rosenberg's conception of it is severely impoverished.

Laubichler and Wagner's general issue is with Rosenberg's habit of using the terms molecular developmental biology and developmental molecular biology interchangeably, whereas for them, these terms refer to two different projects with explanatory agendas of their own (Laubichler and Wagner 2001, pp. 56–57). The former is in the business of studying developmental phenomena, and employs the tools of molecular biology to this end. The latter project is in the business of studying biologically interesting molecules that figure in developmental

processes, and asks questions concerning these molecules and their interactions. In the former, the explananda concerns any phenomena that can be addressed from a developmental perspective. In the latter, the explanatory agenda is by default set as whatever questions can be addressed by studying the properties of molecules. According to Laubichler and Wagner, Rosenberg claims that his arguments bear on the former project, but then really concerns only the explanatory aims of the latter, thus securing his preferred conclusion of molecular reductionism a priori. Laubichler and Wagner claim that the latter perspective is not sufficient for understanding the role of molecular processes in answering properly developmental questions.

As the key evidence from actual biological practice, Laubichler and Wagner consider the phenomena often encountered in developmental biology, where the same developmental outcome has many sufficient determinants, or the same component is associated with many different developmental outcomes. Consider the first, many-one relationship first. Genetic redundancy, for instance, might make it impossible to assign explanatory responsibility over a developmental outcome to any single dedicated molecular component (Laubichler and Wagner 2001, p. 60). In the case of redundancy, a type of pre-emption situation obtains, where many functionally similar genes could bring about a developmental outcome, but only one is actually recruited to do so in the developmental process. Treating a single gene as *the* molecular cause responsible for development would be a mistake, the outcome would have occurred even if this gene were deleted. Another example offered is the function of *bicoid* in the development of drosophila anterior-posterior axis. Maternally transcribed bicoid RNA controls the formation of anterior-posterior axis in drosophila, yet many other species in which bicoid does not exist have a similar body-plan (Laubichler and Wagner 2001, pp. 65–66).

Clearly, if some developmental capacity of an organism is buffered against mutation by genetic redundancy, then none of the redundant genes is individually a difference-maker for the capacity. Given the difference-making analysis of explanation, none of the individual genes alone therefore explain the outcome. But this is not a problem for the reductionist. Redundancy is akin to causal overdetermination in etiological explanation, where multiple causes simultaneously tend for the same effect. The redundant genes are all components of the organism, and would support manipulations of its developmental capacities when jointly manipulated. The direction of explanation still flows from the molecular components up to the system – knowledge of these genetic components is needed in order to manipulate the developmental phenomenon they are associated with.

Similar story goes for the bicoid-case. Bicoid is part of a constitutive explanation of a developmental capacity in drosophila, the capacity to form a body-plan organized along the anterior-posterior axis. Observing that other species, too, develop similar structures is observing the manifestation of a similar capacity in structurally different systems. In each case, the explanation of this developmental capacity amounts to showing how it is constituted. In other species than drosophila, different types of components might explain a similar capacity. But in each case, the explanation flows from the (molecular) components to the capacity of a system.

The fact that the constitutive basis is different in different species is akin to the fact that there can be varying etiological causes that lead to similar effects. It is of course false to claim that bicoid is the unique cause of axis-formation in every case, but given a little charity, this probably isn't what Rosenberg is saying. Rosenberg is right if his claim is merely that in every case of this type, the explanation requires a bottom-up –account of how the organism possesses a capacity to form a body-plan. Devising such explanations seems to be what Rosenberg's developmental molecular biology would be about.

But addressing the constitutive questions does not exhaust the explanatory agenda of developmental biology. Recall that pure constitutive explanation starts with a question of why a system possesses a capacity Z rather than Z′, and answers this by reference to the system being made of components O rather than some other components or different configuration of components O′. The components invoked in the explanation, in turn, must come from somewhere, and must have been organized by some causal process, and one can ask how this happened. To address this question, one takes the contrast between O and O′ as an explanandum of its own. This perspective is lacking from Rosenberg's construal of developmental biology, and this I take is what Wagner and Laubichler take issue with. Only by taking such a developmental perspective it is possible to make sense of Wagner and Laubichler's other class of examples, that of a same molecular component leading to different outcomes within the same organism.

Wagner and Laubichler mention various examples of this kind, but here it suffices to discuss one to make the point. The molecular pathway involving genes *wingless* and *hedgehog* and their protein products, and the enzyme zest-white 3 kinase "is involved in the formation of segment boundaries in *Drosophila*, but also specifies the proximodistal axis in the eye, leg, and wing imaginal disc" (Laubichler and Wagner 2001, p. 63). That is, the same molecular structure is constitutive of different developmental capacities in different situations.

Sameness of constitution resulting in different systemic outcomes seems like a violation of supervenience. Since this cannot be the case, the change in the role of *wingless/hedgehog*-pathway becomes an explandum to be addressed in itself. Here, we have an explicitly contrastive question: why is the pathway associated with an outcome Z in a developmental situation A, and outcome Z′ in a situation B? The answer lies in the differences in the causal context of the pathway in these situations. This context, in turn, consists of other components of the system, and changes over the course of development.

In order to explain changes in the developmental role of the pathway, we must invoke developmental capacities that manifest as changes in the system's components and their organization so that the causal context of the *wingless/hedgehog*-pathway within the organism changes over time. If these capacities could have resulted from other configurations of molecular components than what is actually observed, then tracking the actual developmental changes in molecular detail might not capture the relevant dependencies. Furthermore, since the full story would include accounting for the manifestations of those capacities, the environment cannot be excluded from the picture, since manifestations of capacities depend on

the right kind of environmental input. Simply following the changes in the system's constituents wouldn't therefore capture all the relevant difference-makers.

The web of dependencies between the constituents of a system and its developmental capacities, and their coupling with the relevant environmental factors is what forms the reference frame that allows biologists to ask questions about how the components of the organism together with environmental factors bear on the outcomes of developmental processes. The more limited explanatory agenda of Rosenberg's developmental molecular biology can address only some of the questions that can arise in such a framework.

7 Conclusions

In this paper I have characterized explanation as tracking change-relating dependencies that afford manipulating the explanandum by intervening on the explanans. This allows us to distinguish between etiological and constitutive explanation, and to characterize the framework of developmental explanation that combines these two. The difference between etiological and constitutive explanation was shown to be about where the contrast in the explanatory request is located. These differences correspond to differences in what the contrast in the explanans is about, and what kind of interventions are appropriate to consider.

Etiological explanation involves considering what would happen to a behavior of a system if external conditions were changed by interventions. Constitutive explanation considers what would happen to the capacities of a system if interventions were to change its components. Etiological explanation treats the constitution of a system and its causal capacities as static while external conditions vary, while constitutive explanation treats the causal environment of a system as static while the system's components and their organization are imagined to vary, while considering how the variation would map to variation in the intended explanandum, respectively.

Developmental explanation is an amalgam of etiological and constitutive explanation. It involves explaining the causal capacities of a system constitutively, as well as explaining changes in the constitution of the system as the causal consequence of the manifestation of the system's capacities. Neither the system's constitution, nor the conditions external to the system can be treated as static within a framework of developmental explanation. However, no actually tractable explanation can explain every aspect of a system at once, therefore actual developmental explanations address questions about specific aspects of the developmental process one at a time.

I have assessed Alex Rosenberg's molecular reductionism about developmental biology as well as a critique of this account due to Laubichler and Wagner using the framework of developmental explanation presented. Rosenberg is partly right about the explanatory status of dispositional notions like the organizer region. The organizer and related notions do not point to targets for intervention that would enable manipulating *specific* developmental capacities of an organism. If such capacities are taken to be the target of explanation of developmental biology,

then the organizer concept and similar concepts explain little. When the target of explanation is a capacity, the explanans should point to the constituents that make a difference with respect to that capacity. The organizer at best hints at such an explanation. Typical explanations of contemporary developmental biology target a specific capacity of an organism, such as the capacity of a fruit-fly to develop a segmented body-plan. Such explanations are reductive in the sense that the explanatory dependency runs bottom-up in a part-whole hierarchy. Laubichler and Wagner's example of many-one relations between molecular components and developmental outcomes does not raise a problem for this kind of explanation; it just means that the scientists must gather information from multiple experimental interventions and combinations of interventions in order to come up with an adequate picture of what components are relevant for the capacity in question.

However, pure constitutive explanations of system-capacities do not exhaust the explanatory agenda of developmental biology. In the framework of developmental explanation, stable configuration of components, that explains properties of a system, is not taken to be a pre-existing fact. Rather, biologists ask questions about the development of the components and their configuration itself. Laubichler and Wagner's other example is of this type: it is about explaining why the constitutive role of some molecular structure changes within the system over time. Explaining this requires reference to factors that are not captured simply by describing the actual molecular components and their interactions over time. This is due to the fact that many different configurations of components might give rise to the same developmental capacities that are responsible for developmental change, as well as the fact that the explanation involves reference to manifestations of developmental capacities and to their environment.

References

Achinstein, P. 1983. The pragmatic character of explanation, In *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association, Volume Two: Symposia and invited papers*, 275–292. Chicago: The University of Chicago Press.

Craver, C. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Cummins, R. 2000. 'How does it work?' versus 'what are the laws?': Two conceptions of psychological explanation. In *Explanation and cognition*, ed. F. Keil and R. Wilson, 117–144. Cambridge: The MIT Press.

Frost-Arnold, G. 2004. How to be an anti-reductionist about developmental biology: Response to Laubichler and Wagner. *Biology and Philosophy* 19: 75–91.

Garfinkel, A. 1981. *Forms of explanation*. New Haven: Yale University Press.

Glennan, S. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69(S3): S342–S353.

Keller, E.F. 1999. Understanding development. *Biology and Philosophy* 14: 321–330.

Laubichler, M., and A. Wagner. 2001. How molecular is molecular developmental biology? A reply to Alex Rosenberg's reductionism redux: Computing the embryo. *Biology and Philosophy* 16: 53–68.

Machamer, P., L. Darden, and C. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67: 1–25.

McManus, F. 2012. Development and mechanistic explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 532–541.

Rosenberg, A. 1997. Reductionism redux: Computing the embryo. *Biology and Philosophy* 12: 445–470.

Rosenberg, A. 2006. *Darwinian reductionism or, how to stop worrying and love molecular biology*. Chicago/London: The University of Chicago Press.

Salmon, W. 1984. Scientific explanation: Three basic conceptions, In *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association, Volume Two: Symposia and invited papers*, 293–305. Chicago: The University of Chicago Press.

Schaffer, J. 2005. Contrastive causation. *Philosophical Review* 114(3): 327–358.

Soto, A.M., C. Sonnenschein, and P.A. Miquel. 2008. On physicalism and downward causation in developmental and cancer biology. *Acta Biotheoretica* 56: 257–274.

Wimsatt, C. 2007. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Woodward, J. 2003. *Making things happen. A theory of causal explanation*. Oxford: Oxford University Press.

Ylikoski, P. 2013. Causal and constitutive explanation compared. *Erkenntnis* 2 Supplement: 277–297.

Ylikoski, P., and J. Kuorikoski. 2010. Dissecting explanatory power. *Philosophical Studies* 148: 201–219.

What Counts as Causation in Physics and Biology?

Jan Faye

1 Introduction

The notion of causation has been with us for a long time. In this paper I argue that the concept of causation has its origin in the biological evolution of higher organisms and that the basic notion of causation is embodied due to our ancestors' interaction with their environment. So I hold that the sense of cause and effect came into the world millions of years before human sapiens developed science and advanced technology. It means that our predecessors primarily had an embodied sensibility of causation based on one or more innate schemata, and only much later acquired a conscious ability to make a reflective application of it. The reflective sense of causation developed whenever human beings began to apply the innate schema of causation to relations completely alienated from the human body. This might have happened in connection with understanding natural powers or alleged supernatural powers, seasonal changes, the movements of the planets, or if we look into the recent history of humankind, in connection with scientifically described phenomena.

A naturalist philosophy has implications for a proper understanding of causation. The notion of causation is a general concept that goes far beyond any particular science. The concept stems from one or more innate schemata of comprehension that are applied to whatever phenomena that fulfil some very common criteria. These criteria are domain-independent in the sense that they do not depend on the realization of a certain type of events. It does not make much sense to claim that the concept of causation has only a physical meaning, since it also helps us to understand other types of events such as particular biological phenomena,

J. Faye (✉)
Department of Media, Cognition, and Communication, University of Copenhagen, Njalsgade 80, 2300 Copenhagen, Denmark
e-mail: faye@hum.ku.dk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_12,

particular economical phenomena, particular cultural phenomena, etc. The concept covers many different kinds of manifestation depending on the context in which it is applied. The only requirement is that causes are that which makes a difference and that this difference results in an effect.

Finally, I shall emphasise that causation should not be identified with explanation. Causation is the ontological counterpart to causal explanation. Nevertheless, I argue that it is our explanatory interests that determine what we consider causation to be as long as what is explained fits some very loose criteria of "causation". Although these criteria are embedded in the innate schema of causation, they are only partly due to the biological organisms' interaction with their environment. The criteria exist in humans as a combination of the organisms' interactional behaviour and their ability to abstract thinking.

2 Causation and Evolution

Studies show that birds and mammals possess an embodied sensibility of causation.[1] We know that birds and mammals have a very sophisticated spatial sense.[2] Some of them also seem to have a developed time sense.[3] I take embodied cognition to be whatever belief or understanding an *individual* organism has acquired through involving its body in an information-gathering activity in its environment. The embodied sense of causation produces a fairly robust grasp of causation that these birds and mammals use in their thinking about the environment. Recent field studies and controlled experiments have demonstrated that cows and ravens can foresee what is going to happen in case they would behave in a certain way. They can act according to certain expectations. They can plan and imagine possible outcomes of their action, and whatever their purposes are they can realize the outcomes that correspond to their *actual* purposes.

Just to give one example. It has been observed that a particular pair of ravens had established a sophisticated hunting technique.[4] The male chased red crossbills down an alley between two buildings while the female interrupted their flight so the crossbills turned their direction and flew into a pane of glass in one of the buildings. Because of the impact with the glass the crossbills would either die immediately or fall to the ground unconsciously. In any case the ravens could then feed on their

[1] See, for instance, Miller and Matute (1996).

[2] See Brown and Cook (2006) and its many references.

[3] Evidence supporting a time sense in higher animals can be found in Raby et al. (2007), Kalenscher and Pennartz (2008), and Stephens (2008).

[4] This example is taken from Marzluff and Angell (2012, pp. 75–76). The book describes the intensive research that has been conducted with ravens, crows and magpies and the astonishing results.

corpses. It should also be noticed that the hunting procedure of the ravens was not accidental. It came about by planning and deliberation. During a summer this couple was able to kill around 250 crossbills in this manner. Indeed there is an element of learning, but learning just helps the individual to acquire knowledge about how to carry out a particular causal task in the actual circumstances.

Analogous examples can be found among higher mammals. Lions chase their prey by acting together, by coordinating their movement with whatever each member of the group does with respect to the preys' direction of escape. They are able to foresee what will happen and adapt their behaviour accordingly and then intercept the movement of their game. Monkeys and primates are likewise able to plane in advance and act so that they reach their goals. Chimpanzees teaming up for hunting monkeys have different causal role to play. There is the driver who chases the monkeys ahead of him, there are the blockers who stop the monkey of escaping to the sides, and finally there is the ambusher, the most experienced chimpanzee hiding in the front. His task is to grab and kill one of the approaching monkeys. This would indeed not be possible unless the chimpanzees did not have a fair grasp of the causal processes in their environment. One should also notice that the chimpanzees' role shifts depending on their hunting experience.

The story of evolution of these embodied capacities seems to be the following: little by little, through variation, selection, and retention, reptiles', birds', and mammals' interaction with their environment developed an innate cognitive schema of causation. These schemata are then instigated in each individual as bodily acquired understanding via learning by doing. Experiencing their own actions of bringing about events is the basic cognitive vehicle that helps them to come to terms with the world. The presence of an innate causal schema gave organisms an ability to acquire particular causal beliefs and make particular causal action. They could not have purposes, intentionally making plans, and carry out successful behaviour without having a concept of causation. Indeed, this holds for human beings as well.

3 The Inherited Criteria of Causation

What has been said so far indicates that higher animals can think about their environment and the means of thinking are concepts. In my opinion many vertebrates possess concepts even though they neither have any advanced language nor have the capacity for self-reflection. I take a concept to be defined as the capacity of distinguishing kinds from individuals. If an organism recognizes an object as belonging to a particular sort, it seems fair to say that it has developed that particular concept. Mastering concepts seem necessary for thinking "in terms of" causation, for having causal thoughts.

Let me explain how. In the beginning, vertebrates had developed a behaviour according to which they automatically interacted with the world without having any form of understanding. Eventually organisms were selected according to their capacity to make a behavioural distinction between actions that were rewarding and

actions that were not. The next steps in the evolution seem to be the selection of organisms that were able to reinforce rewarding actions by being able to classify them and to orient them towards a goal. As Hume already observed, any causal process is a series of contiguous events. In order to see the series of events as causal, actions and goals have also to be regular; that is the same type of events has to succeed the same type of events. This is exactly what we witness with respect to the ravens. The same set of behaviours is brought about because this are expected to be succeeded by the same set of events.

But this is not all. An analysis of the various steps in the ravens' behaviour reveals that their embodied conception of causation consists of more than a simple "mirroring" of regular succession among types. Their causal schema mirrors the following elements: (1) a specific type of action brings about a certain type of effect in the proper circumstances, i.e., a specific type of action is effective only with respect to a particular type of environment; (2) a certain type of action prevents a certain type of effect to occur in the proper circumstances; and (3) certain types of events have causal priority to other types of events. The ravens' particular actions were causally successful only in this particular type of context in which there was a corridor between two buildings and one of them had a huge window pane at the end of the corridor. The male's chase of the crossbills produced the crossbills' attempt to escape down the alley towards the window, and the females' blocking their escape route made the crossbills turn left into the window pane instead of right into the free. The action of the male and the female were not only necessary but jointly sufficient for the killing the crossbills. The ravens had learned which events were causally dependent on which. They had learned to bring about a wanted effect.

If these ravens did not possess such an advanced concept of causation, we could not explain the ravens' ability to act strategically by making plans according to their intentions and carrying out these plans. They could foresee what was sufficient for them to do in order to reach a certain goal. They could imagine the possibilities in advanced and by action they could realize the possibilities they wanted to happen.

Indeed, an organism like a raven may sometimes experience an accidental series of events which is not a causal series. So how did birds and mammals in the first place learn to distinguish between accidental regularities and non-accidental regularities? They had to understand causal matters by developing a mental capacity of abstraction and construction.[5] It was necessary for them to acquire the ability to distinguish what in their environment is causally dependent of what, which seems to require the development of a sense of modality. First an organism had to gain the capacity of subtracting features, and then it had to obtain the ability of adding features in order to reach a robust concept of causation. The way of abstraction is carried out by cognitively removing those features of their behaviour and the

[5]I discuss this issue further in Faye (2010).

environment that tie them to the actual circumstances. Having acquired the ability of subtracting features from actual actions and events, an organism got to the notion of sorts, i.e. it was able to identify its actual action and its actual effect to be of a similar type as previous actions and effects.

But still this is not enough for the organism to distinguish between accidental and non-accidental regularities and thereby to know what is causally dependent on what. In addition, the organism needed to gain the ability to add further features to the actual events which would make a difference between them in possible but not actualized situations. An organism must be able to remember situations where its actions were successful in producing a wanted effect and compare them to situations in which its actions were not successful. So in order to develop a useful concept of causation, an organism had to be cognitively able to remove features of the earlier circumstances, which were not causally relevant, and be able to imagine how the circumstances have to be if it should be able to carry out similar actions in the future. For an organism its action would appear non-accidentally related to an effect, i.e., causally related, if and only if it possesses information that its action in similar circumstances could bring about the wanted effect and information that abstaining from acting would not produce the wanted effect.

Concrete actions and individual events exist as such in space and time, actual actions and actual events exist here and now, thus actions and events similar to the actual ones exist at other spaces and times than here and now. However, causal dependency implies more than mere actual succession and contiguity of action and its outcome. Nobody can directly observe in any particular case what the modal features attached to the embodied schema of causal dependency are. No organism can experience that an action can bring about a certain effect before they occur. The modal feature that a possible action *could* bring about a certain kind of effect in case the circumstances appropriate is not empirically accessible in any immediate way. It has been constructed by induction from what an organism remembers about similar actions in other situations. The organism makes up the modal feature in virtue of having experienced what happened in relevant but different circumstances in which it controlled similar events or intervened in their succession. Thus, the abstracted and constructed features as they were disclosed to animals through control and manipulation with their environment became the modal features which we humans identify with the understanding of causes, because even though the non-accidental patterns were observed for past events, they are generalized to apply to any present or future cases of causation under similar circumstances.

4 From Embodiment to Reflection

Causes are differences that make a difference. The embodied conception of causation are functional. The criteria for an organism to think in causal terms are functional criteria. A certain kind of bodily movements can act as causes of

a certain kind of effects if, by carrying out an action of this kind in the proper circumstances, the organism is usually successful in reaching such an effect as its intended goal. There goes, as we just saw, a direct line from the functional criteria of causation to a modal characterization of causation. It is by bodily manipulation and intervention together with mental abstraction and construction that organisms have formed a useful concept of causation. It is by the very same criteria, which led the cognitive evolution to establish such an innate schema of causation as a cognitive means of understanding, that we determine whether the concept of causation applies to processes which are alienated to our bodily actions or perceptually unobservable. Sometimes in the past – after our predecessors had developed a reflective consciousness – the notion of causation already installed was increasingly applied to other things than actions and our immediate environment. It was extended to cover all kinds of things in order to gain understanding in form of experiential coherence and predictability.

Thus, the reflective consciousness automatically wants to see happenings in the world in causal terms because this form of comprehension is evolutionary induced in our cognitive apparatus as one of the basic schemata of understanding. We cannot but describe causation to a particular connection of events if we can ascertain that the events satisfy the criteria of being causes and effects.

When people appeal to their causal intuitions it is, I think, this embodied understanding of causation (grounded in an innate causal schema) they are hinting at. But philosophical reflection and deliberation also add some features to our conception of causation: consider once again the ravens' hunt of the crossbills but from our reflective perspective. The kind of action that a raven uses to bring about a certain outcome is regarded by us as what necessitates the effect in the circumstances, and the action whose existence does not depend on the outcome is claimed to be necessary for the effect in the circumstances. Let us call an actual set of actions for *c* and the hitting-the window pane of a particular crossbill for *e*. We take *c* to *necessitate e* in a strong sense according to which this particular *e* had not occurred under the actual circumstances in case the ravens would have abstained from doing *c*. This belief is due to our observation that under the similar circumstances these ravens reached their goal whenever they performed a type of action like *c*. But we also take *c* to be *necessary* for *e* in the sense that if *c* had not occurred, then *e* would not have occurred in the actual circumstances. Indeed, this is not something we can observe. It is a modal feature we have constructed from observation of similar cases where an event of a similar kind like *e* did not occur without an action of a similar kind like *c* occurred. We could have shot the ravens and no crossbill would have hit the pane of glass. It is these observations of similar but numerically different cases that eventually have become embedded in our reflective notion of causation as the modal features of an actual non-accidental succession of events. Hence our readiness to ascribe counterfactual implication to causal beliefs as part of our reflective understanding rises from our experience with similar types of events other than the actual events under consideration.

5 Causation in Physics

There would have been no science without a reflective consciousness and an advanced language for communication. Understanding nature by science builds nonetheless on our biologically inherited notion of causation. Therefore there is no particular way in which either causes or causal processes are constituted in science. Causal relations do not have some specific nature. What counts as a cause or a causal process depends on the vocabulary of the particular science and the phenomena we want to study. As long as a certain set of phenomena obeys a relationship that can be determined by our inherited criteria of causation it counts as a causal relationship. At the same time it must be granted that science has developed advanced methods of finding causal relations which involve statistics and probability measures.

In physics we find several candidates for being a causal process. Four different suggestions are usually put forward: (1) the causal link can be identified with the transference of (positive) energy; (2) it can be identified with the conservation of physical quantities like charge and linear or angular momentum; (3) it can be identified with the interaction of forces; or (4) it can be identified with microscopic interaction in the framework of quantum field theory. Even though I have defended (1) in the attempt to understand backwards causation (Faye 1989), I still hold that none of these four identifications can be said to characterize the true nature of causation in physics (Faye 1994). Which one is basically more to the point than the others cannot be determined independently of the context in which the discussion takes place. Moreover, the four suggestions do not even exhaust the possible kinds of causation in physics. Rather they blur a distinction between causes and causal processes which in many contexts makes sense. A cause may initiate or trigger a causal process but it is not in itself a part of the process it starts. Again an effect is an interacting cause that acts as the determinator of the process.

Hence I want to suggest the following distinction between causes and causal processes. *Causes* are always seen with respect to a system and its surroundings; they appear each time a system affects its environment (another system) or the environment affects the system, whereas *causal processes* are those activities that uphold the existence of a given system. Causes give rise to *causal laws* if they always occur under regular circumstances, whereas causal processes give rise to *causal mechanism* if they occur under regular circumstances. I think such a distinction by and large reflects the implicit use we make of these terms.

The following example from quantum mechanics illustrates my claim. When an electron, say, in a hydrogen atom is in an excited state there is a non-vanishing probability that it will transit from a higher energy level to a lower one within a very short interval. While transitioning the electron emits a photon whose wavelength depends on the difference between the two energy levels. It starts a process. So it seems reasonable to say that it is the transition of the electron that causes the atom to emit a photon with a specific frequency. But none of the four proposals describe the actual transition and therefore the actual cause. Yet, it is still a causal law that

electrons "jumping" from a higher to a lower state emit photons and that they have a definite wavelength depending on the principal quantum numbers.

Another example is taken from classical mechanics. The moon moves around the earth, and according to Newton's theory of gravitation this is due to the gravitational force of the earth. Without the presence of the gravitational force the moon would move in a straight line. Thus in this context it makes sense to say that the gravitation of the earth causes the moon to move around the earth because it is the gravitational force that makes a difference. It is the total mass of the earth that has the ability to move the moon around in elliptic circles, and it is the moon that would move uniformly in a straight line if it was not been attracted by the mass of the earth. Now it is, I think, questionable to reduce the mechanical description of the movement of the moon to a process of interaction. Processes may change over time but changes are what causes do. Try to think of an attempt to understand the movement of the moon in terms of microscopic interaction. This form of interaction consists in the exchange of so-called intermediary particles among other fundamental particles. The four basic forms of interaction are gravitational, electromagnetic, the weak and the strong interaction. The intermediary particle of the gravitational force is suggested to be the massless graviton. If this hypothesis is true – the graviton has yet to be discovered – it means of course that the gravitation of the earth could not cause the movement of the moon unless the fundamental particles that make up the earth and the moon exchange gravitons. But it is not easy to see how the exchange of gravitons could make a difference. The exchange of gravitons is a process which does not change anything as long as it continues to take place. And if the exchange of gravitons changes its exchange rate there must be something external which causes the changes.

What the two examples have in common is very minimal. Transition in an atom and gravitational force are different kind of things. But they count as causes because they both make a difference. Thus, I agree with Nancy Cartwright (1999) when she claims in *The Dappled World*: "there is a great variety of different kinds of causes and that even causes of the same kind can operate in different ways." (p. 104); and she, as well as I, have learned from Elisabeth Anscombe (1971) that "the word 'cause' itself is highly general … I mean: the word 'cause' can be *added to* a language in which are already represented many causal concepts" (p. 68). The problem is not that various philosophical accounts of causation do not add up in their own right but that there exists a hegemonic tendency among supporters of extending particular theories of causation to all phenomena. However, I am a bit more hesitant when Cartwright (2002) in a later paper maintains that

> The problem is not that there are no such things as causal laws; the world is rife with them. The problem is rather that there is no single thing of much detail that they all have in common, something they share that makes them all *causal* laws. These investigations support a two-fold conclusion 1) There is a variety of different kinds of causal laws that operate in a variety of different ways and a variety of different kinds of causal questions that we can ask; 2) Each of these can have its own characteristic markers; but there are no interesting features that they all share in common.

I have no problems with the first part of the first part of the quotation, and I also agree with conclusion (1). The conclusion (2) and the second part of the first part are those claims about which I am more uncertain. The above examples show that our understanding of causal laws has some functional features in common. Perhaps not many – but still enough to protect the use of causal terms from being accidental. Both the transition of an electron and the gravitational force of the earth have in common that they bring about changes, and their discovery helps us to at least understand nature. The use of causal terms obeys certain minimal criteria which imply that we are ready to commit ourselves to a discourse of counterfactuals. Materially, causal laws may be very different kinds, but functionally they are very alike.

6 Causation in Biology

Unsurprisingly, what counts as causation in biology also varies with context. What kind of analysis biologists find explanatory depends on epistemic and pragmatic considerations. The scientist chooses the form of explanation depending on her research interests and then finds the causal relationship that corresponds to these interests, i.e., she finds the relationship among the phenomena she wants to study that fulfills the criteria of causation and serves her research interests. These interests may concern generic, evolutionary, physiological, neural or behavioral factors. For some analyses the level of genes will indeed tell us many things about the causes that some scientists are interested in. But this does not imply that an analysis on this level alone would be sufficient for understanding more complex causal relations. Other scientists believe that organisms are complex self-organizing systems that can be attributed "agency", functional part-whole relations, historicity, etc. These scientists argue, correctly I believe, that much about organisms and their environment has to be understood on this level. Such an approach indicates an organismic or mereological point of view. I agree with Dupré that on the *ontological* level methodological reductionism does not stand strong (Dupré 2007). It is merely one out of many fruitful approaches but not "the only game in town". So let us dig a bit deeper and illustrate the last claim with some examples.

In recent years debates about explanation and causation in biology have been related to questions regarding descriptions of *mechanisms* as the prototype of modern biological explanations. The so-called New Mechanistic Philosophy is a reformulation of biological explanations in terms of mechanisms, but without the reductionist tendencies associated with earlier accounts of mechanism. An important contribution is the MDC model which was formulated by Machamer et al. (2000). Their central suggestion is:

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions (p. 3).

According to the MDC-model, entities and activities are interdependent; thus, the model does not focus only on properties of entities nor only on activities and processes. Causation is a result of entities carrying out their activities. This characterization of causation in biology stands in stark contrast to earlier formulations in terms of causal laws. One such formulation is due to Glennan who in an earlier work defined a mechanism as "a complex system which produces that behavior by virtue of the interaction of a number of parts according to direct causal laws" (Glennan 1996). But Machamer, Darden and Craver point out that one rarely finds such "direct causal laws" in biology, and even when generalizations are stated, the intelligibility of mechanisms is not directly reducible to their regularity but involves higher-level productive activities in inter-level explanatory models (Machamer et al. 2000, pp. 21–22).

Nobody would deny that the MDC-model provides some rewarding insights into causation in biology but it is doubtful whether it can cover all forms of causation from genes to organelles to cells to tissues to organs to whole organisms. The problem is at least twofold: Are there always definite entities involved whenever biologists point to a cause of a specific effect? Are the activities always of a kind that can be accounted for by inter-level descriptions?

Concerning the first question, Dupré (2007) for one, has argued that many biological entities such as genes are not stable enough to carry a mechanism. On the ontological level, genes do not exist as persistent material entities. A given protein cannot be equated with a fixed part of a DNA-sequence, and therefore genes are now in many contexts functionally rather than structurally defined. The difficulties of establishing clear identity criteria of entities challenge a view of causation that focuses, even only partly, on properties of entities. Furthermore, finding the causally relevant entities that can be said to carry out activities also turns out to more complicated than previously thought.

Just as important, I think, is the fact that the MDC-model does not clearly distinguish between causes and causal processes. On the explanatory level the distinction between causes and causal processes (mechanisms) seems relevant. An appeal to a cause often constitutes the appropriate respond to possible *why*-questions, whereas a reference to a causal process does not answer any particular why-question but rather a series of possible *how*-questions. From a conceptual and ontological point of view the distinction also allows us to introduce other explanatory approaches. Hans Reichenbach talked about processes, like a beam of light, as a series of genidentical events that are supported by the same objects that do not change their properties over time. In contrast, the MDC-model sees the process as being "productive of regular changes", and I take this to mean that these changes are not merely spatial replacements. But this conception implies that there is no room for a clear distinction between a process and a cause that changes the process. One way to state a difference is to say that a cause is always external to the process; it belongs to the environment of the process so that any change of a particular process (apart from its spatial replacement) is caused by some factors external to the process itself. Indeed such a clear distinction is sometimes important as when we want to understand the causal interaction, say, between an organism and its environment as

in evolutionary biology, etiology, and theories of infections, or between a specific process and the changes that may happened to such a process caused by other processes external to it. It is within a discourse of causes as separated from causal processes that it makes sense to talk about biological laws. I am not thereby claiming that the goal for every scientific investigation is to find causal laws, but saying that sometimes this kind of ceteris paribus generalization offers valuable understanding of regular changes of those objects it covers. Moreover, it seems that much of our knowledge of organism and their behaviour stems from finding causal regularities.

For years it has been assumed that genes were the cause of the phylogenetic features but the problem which this assumption is, for instance, that the same genes are present in all cells of an organism even though these cells develop very differently in relation to their position in the organism. One response to this conundrum is the Developmental Systems Theory (DST). This view claims that genes do not have causal priority and no singular relationship exists between genes and the phylogenetic development. Rather heredity and development are equally caused by genetic, environmental, and epigenetic factors. Thus this approach holds that there are many equally important factors to take into consideration when analyzing developmental processes instead of focusing on genes alone (Sterelny and Griffiths 1999). We must take the whole developmental matrix into consideration. Genes are just one out of many resources. Much more than genes are inherited from one generation to the next; these elements include membranes, organelles, methylation patterns, cellular chemistry, and behavioral patterns. Thus, to understand inheritance and development, one must take the whole system including the environment into account.

A related but even more general view is Systems biology (SB). It is a new approach in the life sciences that emphasizes the non-linear relationships between biological variables and the importance of *organization* for understanding living systems. A central method in SB is large scale mathematical and computational modeling paired with an increased focus on quantitative analysis. Robert Rosen, a central figure in Systems biology, suggests that instead of focusing on material causation of behavior (properties of entities) we should start looking for principles that govern the *organization* of phenomena (Rosen 2000; Wolkenhauer 2001). He defined the new approach as *relational biology* which emphasizes the need to transcend the reactive paradigm of machine analogies and understand biological organism quite independently of their physical or chemical constitution (Wolkenhauer 2001). So systems biology attempts to move away from the biological entities and concentrates only on the dynamical process. In Rosen's view the way to progress in our understanding of biological systems is to choose *functions* and not *structures* as units of analysis. Systems biology marks a reaction to earlier reductive tendencies in biology and one of its aims is to discover and describe the function of emergent properties of the systems.

The SB model assumes that the properties of any level depend both on the properties of the parts "beneath" them and the properties of the whole into which they are assembled. Thereby the SB-model introduces emergent properties of the system. But the obvious question is then how these levels interact with one another;

i.e., whether we have efficient downward causation from the whole to the parts. It is difficult to see how a notion like downward causation can fulfil the criteria of causation. We may bring about the whole by putting the elements together, but it seems to be impossible to bring about the elements by creating the whole. The whole cannot exist unless its elements already exist. A related problem seems to be that we think of an event as acting as a cause only with respect to given circumstances. But under which conditions does the whole act as a cause? Avoiding such problems might be to argue for only a "medium" version of downward causation, where higher level properties are constraining conditions for the activity of lower levels, but there is no efficient causation from higher to lower levels (Emmeche et al. 2000). In my opinion such a view on the part-whole relationship explains why biologists in many cases resort to functional descriptions where the actual effect is used to account for the function of the cause.

Looking at causality at the level of human beings one often meets neuro-physical reductionism, which equips the brain with properties that only make sense in consideration of the whole living organism. However, it does not make sense to claim that the brain makes decisions and then look for a place in the brain where the ability of decision making is placed. If somebody nevertheless claims so, he or she commits what has been called the mereological fallacy. We cannot identify an activity in the brain with making a certain decision unless we already have a clear concept of what a decision is independently of any neurological structure. A decision is a mental commitment to one of several possibilities of action after a person has deliberated which of them he or she should pursue. It is human beings, who make decisions, and indeed we could not make them without our brain, but this does not imply that it is the brain that makes decisions. When scientists want to understand *Homo sapiens* (and other higher animals) as a psychologically and socially influenced creature they look for forms of causal impacts which exist at the level of societal and cultural induced beliefs. Here we take recourse to intentional explanation where we understand behavior and agency according to biologically innate intentions.

7 Causation and Emergency

Indeed, it is not only within biology that we may find emergent properties. In physics we see them, too. Both biologists and physicists see them if they wish. There is a stride in physics between particle physicists and condensed matter physicists about reductionism and emergentism. The latter believe that the reach for an ultimate theory in terms of elementary particles and fields is essentially mistaken since there is no explanatory link between such a hypothetical theory and most real physical phenomena. These phenomena are emergent in the sense that they depend on organizing principles and collective states that cannot be reduced to simpler states of elementary particles. As a response to the strong reductivist and formalistic attitude behind finding a theory of everything, Robert Laughlin and David Pines, the first a

Nobel laureate of physics, turn against the "imperative of reductionism", saying that it "requires us never to use experiment, as its objective is to construct a deductive path from the ultimate equations to the experiment without cheating." And they conclude:

> Rather than a Theory of Everything we appear to face a hierarchy of Theories of Things, each emerging from its parents and evolving into its children as the energy scale is lowered. ... The central task of theoretical physics in our time is no longer to write down the ultimate equations but rather to catalogue and understand emergent behaviour in its many guises, including potentially life itself.[6]

According to them, the list of emergent behaviours is endless and they mention examples such as the work of proteins, the superconductivity of magnetic insulators, and the superfluidity of ^{3}He. Other physicists, like Joseph Polchinski, deny the existence of emergent properties and hold that "the history of science seems to be a steady progression toward simpler and more unified laws. ... Things may take many surprising twists and turns but we reductionists are still quite happily and busily reducing."

Now let us take the emergence of different levels for granted. There exist properties at each level which cannot be explained by properties that exist below that level. The problem is then the causal status of emergent properties. It is quite obvious that at each level we explain the behavior of a system by referring to the causal behaviour of another system at the same level. The impact of the ball breaks the window. It is also clear that we sometimes can explain the behaviour of a system at one level by reducing this behaviour to the behaviour of the subsystems at a lower level. This is the case when we explain temperature in terms of the kinetic gas theory and bridge principles. However, if emergent properties exist, as we assume, but they do not have any causal effect on those subsystems on which they supervene, what role do they then play with respect to these subsystems? To answer this question we may consider a flock of starlings gathering during the autumn.

The huge flocks may count millions of individuals, and it seems that each individual reacts to the flocking behaviour whenever the entire flock of starlings makes a sudden turn in the air, spreads out, or contracts into a dense ball in rapidly changing patterns. One might think that the emergent flocking behaviour, consisting of density, velocity, flight direction, the change of direction, and the shape of the flock, causally influences the flying behaviour of each single bird. Indeed, based on these properties one might be able to construct a complex differential equation that describes the flock as one system which can be used to explains *how* the movement of the system takes place, but such an equation cannot be used to explain *why* it takes place in each individual. However, focusing on individual starlings scientists have been able to create a causal model which can be used to simulate such flocking behaviour and therefore give the same result as if the behaviour of the entire flock had a direct causal influence on each individual. Basic models of flocking behavior

[6]The quotations can be found in Kragh (2011, pp. 274–275) and this particular one is from Laughlin and Pines (2000).

are controlled by three simple dispositions: *Separation* – avoid crowding neighbors (short range repulsion); *Alignment* – steer towards average heading of neighbors; *Cohesion* – steer towards average position of neighbours (long range attraction). Based on these three simple dispositions, the simulated flock moves in an extremely realistic way, creating complex motion and interaction that would be extremely hard to create otherwise. Two such basic models are available: one is called the metric distance model; the other the topological distance model. The first operates with a focal individual that pays attention to all of the individuals within a certain distance; the second operates with a focal individual that pays attention to six or seven individuals closest to itself. Scientists have established that starlings act according to the second model.

Now, what is interesting is that we cannot causally describe flocking behaviour and *why* it takes place, unless we presuppose the existence of certain dispositions that we attribute to the subsystems but only to the subsystems as members of a bigger system. Evolutionary selection explains why separation, alignment and cohesion exist, but the flocking behaviour explains the function of these dispositions.[7] These shared dispositions can explain *how* the flock behaves but not *why* the individual starling behaves as it does as a member of a huge flock. For instance, a single starling or two mating starlings would not behave according to these dispositions. Instead, the function of each single starling's dispositions can only be understood in the relation to the flock behaviour. Indeed flocking also have a function in relation to an even larger system that also includes birds of prey. The conclusion seems to be, if we can generalize the present example to other cases of emergence, that causal processes within a system determine the emergent features of the system. Nevertheless, these overall effects can be causally explained by the action of the subsystems only in case we can attribute certain behavioural dispositions to the subsystems in relation to their membership of a larger system. It makes sense to ascribe dispositions such as separation, alignment, and cohesion to single birds only if we already know their potential memberships of a flock. It is such an ascription of functional features already to the subsystems, i.e., features they only have as a member of a larger system, which help us to identify the causal processes within this larger system and to describe how these give rise to the emergent behaviour of the entire system. Emergent properties function as causal constraints for lower level behaviour.

8 One System or Many Systems

Quite generally it seems that the lower a level of organization scientists consider with respect to a biological system, the closer they get to the same forms of causation that are considered in chemistry and physics. I don't think that this is due to a single

[7]The role of functional explanation is discussed in my forthcoming book *Understanding by Science.*

ontological or a single epistemological factor. In case scientists study molecular biology the phenomena are not that different from the phenomena they study in chemistry. Here they are interested in understanding phenomena according to the chemical and physical concepts and causal models just as much as their biological functions.[8] Whereas if scientists analyse higher level relations between organs and organism, they tend to understand these relations in functional terms, and whenever they attempt to understand animals looking for food, making decisions, the closer they get to forms of causality in terms of which we understand intentional human being.

However, a distinction between a system and its subsystems is an analytic tool based on certain objective delineations. Regardless of whether we talk physics or biology you may regard spatially separated phenomena as constituting different subsystems of one system or constituting many systems. Take, for instance, the Milky Way and the Andromeda Galaxy. Just by the fact that these two phenomena have different proper names show that they are considered to be two physical systems. This way of thinking is quite reasonable for many astronomical purposes. All the stars in the Milky Way rotate around Sagittarius A*, a supermassive black hole at the center of our galaxy. Similarly the stars of the Andromeda Galaxy are rotating around a supermassive black hole. Many of the physical features of these two galaxies are separated from one another. But we may also consider them as one system, a binary system; because these two galaxies are still so close that they orbit around a common center of mass. The result is that the Milky Way and the Andromeda Galaxy approach one another and become one single elliptical galaxy 3–4 billion years from now. Hence it depends on the research context in which we see them, whether we take them to be two separated systems or one big system interlocked by gravitation.

In general, the same holds in biology. The complexity of the systems studied by biology is much greater than in physics because it has many more degrees of freedom than a physical system. But this does not refute the claim that the same biological phenomenon can both be considered as a system or as a subsystem depending on what it is scientists want to know. To illustrate this point imagine a cell. A cell is often considered as the fundamental building block of life but it consists of many different organelles that can be studied as single systems. However, we cannot understand their function unless we bring in their relations to the cell. In the same way, we cannot understand the function of the cell without understanding its relation to the organ it is part of.

Thus, what counts as a system and what counts as its environment is flexible in the sense that it depends partly on our research interest. A dynamical system, I suggest, is a structure consisting of some stuff that partakes in constant processes that actively keep the system running very much independently of what happens

[8]Marcel Weber points out to me that molecular biologists also take a functional perspective on molecules (in the sense of biological function), while chemists don't. I agree, which I hope the present sentence shows.

outside the system. These internal processes are mostly "indifferent" to the external processes taking place in the environment in the sense that little interaction happens between the internal processes and the external world. However, in certain research contexts one may therefore focus on these processes rather than the environment as long as it makes sense to avoid consider causes that intervene with the processes in question. For instance, the Milky Way is "isolated" from the rest of the universe since we can describe many dynamical processes which run our galaxy without having to consider the gravitational pull from Andromeda or any other galaxy. Nonetheless, it also seems to be the case that we often can decide to classify a system so that one can search for causal mechanics within a system but causal laws between systems (between a system and its environment).

9 Conclusion

Sometimes scientists focus on causal processes, sometimes they focus on causal laws. Ontologically, there is a difference, but when it comes to the practice of explanations both approaches seem to provide complementary perspectives. Whatever level they target, scientists may attempt to understand what constitutes a process (reduction) or they may attempt to understand what external causes the change of process (non-reduction). The latter requires that they bring in the environmental context in which the process evolves. Because of some innate criteria of causation I think it makes little sense to talk about downward causation as efficient causal relations from higher level to lower level. Instead emergent features act as constraining higher level conditions on lower level activities. Thus, I think it makes much sense to claim that there are intra-level causations as well as to claim that the whole may constraint individual causal processes.

References

Anscombe, E. 1971. Causality and determination. Reprinted in *Causation and conditionals*, ed. E. Sosa, 63–81. Oxford: Oxford University Press, 1975.

Brown, M.F., and R.C. Cook (eds.). 2006. *Animal spatial cognition: Comparative, neural and computational approaches.* Available: www.pigeon.psy.tufts.edu/asc/.

Cartwright, N. 1999. *The dappled world. A study of the boundaries of science*. Oxford: Oxford University Press.

Cartwright, N. 2002. Causation: One word, many things. Technical Report 07/03, Centre for Philosophy of Natural and Social Science. Available: www.lse.ac.uk/CPNSS/pdf/DP_withCover_Causality/CTR07-03-C.pdf.

Dupré, J. 2007. Is biology reducible to the laws of physics? *American Scientist – LA English* 95(3): 274.

Emmeche, C., S. Køppe, and F. Stjernfeldt. 2000. Levels, emergence, and three versions of downward causation. In *Downward causation. Minds, bodies and matter*, ed. P.B. Andersen, C. Emmeche, N.O. Finnemann, and P.V. Christiansen, 13–33. Aarhus: Aarhus University Press.

Faye, J. 1989. *The reality of the future*. Odense: Odense University Press.
Faye, J. 1994. Causal beliefs and their justification. In *Logic and causal reasoning*, ed. J. Faye, U. Scheffler, and M. Urchs, 141–168. Berlin: Akademie Verlag.
Faye, J. 2010. Causality, contiguity, and construction. *Organon F: Journal of Analytic Philosophy* 17: 443–460.
Glennan, S. 1996. Mechanisms and the nature of causation. *Erkenntnis* 44: 49–71.
Kalenscher, T., and C.M.A. Pennartz. 2008. Is a bird in the hand worth two in the future? The neuroeconomics of intertemporal decision-making. *Progress in Neurobiology* 84: 284–315.
Kragh, H. 2011. *Higher speculations. Grand theories and failed revolutions in physics and cosmology*. Oxford: Oxford University Press.
Laughlin, R.B., and D. Pines. 2000. The theory of everything. *Proceeding of the National Academy of Sciences* 97: 28–31.
Machamer, P., L. Darden, and C. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67: 1–25.
Marzluff, J., and T. Angell. 2012. *Gifts of the crow*. New York: Free Press.
Miller, R.R., and H. Matute. 1996. Animal analogues of causal judgment. *The Psychology of Learning and Motivation* 34: 133–166.
Raby, C.R., D.M. Alexis, A. Dickinson, and N.S. Clayton. 2007. Planning for the future by western scrub-jays. *Nature* 445: 919–921.
Rosen, R. 2000. *Essays on life itself*. New York: Columbia University Press.
Stephens, D.W. 2008. Decision ecology: Foraging and the ecology of animal decision making. *Cognitive, Affective, & Behavioral Neuroscience* 8: 475–484.
Sterelny, K., and P.E. Griffiths. 1999. *Sex and death. An introduction to philosophy of biology*. Chicago: University of Chicago Press.
Wolkenhauer, O. 2001. Systems biology: The reincarnation of systems theory applied in biology. *Briefings in Bioinformatics* 2(3): 258–270.

Challenges to Characterizing the Notion of Causation Across Disciplinary Boundaries: Comment on Faye

Jan Baedke

1 Introduction

In the article "What Counts as Causation in Physics and Biology?" (this volume) Jan Faye argues for the possibility of characterizing the notion of causation more precisely across disciplinary boundaries and even across the boundary of science itself. The paper has two central points: first, Faye claims that the way we think about causation and the way we causally explain are determined by a general cognitive schema. This schema includes a so-called "embodied sense of causation" – a pre-human phylogenetic "gift" – and a "reflective sense of causation". The latter was developed as an extension of the embodied sense of causation in order to trace those dependencies in nature that are completely detached from the human body. Second, based on this cognitive schema, Faye claims that a metadisciplinary notion of causation can be characterized, which suitably captures what counts as a causal relationship in different scientific explanatory practices.

The idea that our understanding of the world in everyday life as well as in scientific investigations is at least in part shaped by evolution has a long history. Especially with respect to physics and biology, looking for a common cognitive schema which tells us how to conceive of the complexity of nature (e.g. in terms of causation) has been emphasized by physicists as well as biologists (see Helmholtz 1896; Mach 1897, pp. 68–69, 1910, p. 235; Vollmer 1985, pp. 63–64). Faye carries on this tradition. He highlights two explanatory approaches directly linked to physicists' and biologists' general notion of causation: (i) the description of causal processes and (ii) the identification of causal factors. A causal factor is understood as a "determinator of a process" that makes a difference (i.e. it brings about a change) in this process.

J. Baedke (✉)
Department of Philosophy I, Ruhr University Bochum, Universitätsstr. 150, 44801 Bochum, Germany
e-mail: Jan.Baedke@rub.de

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_13,

In this paper, I will address some problems in Faye's approach to a transdisciplinary notion of causation. In Sect. 2, I will highlight a general reasoning strategy – considered only briefly by Faye – which is highly important for answering what counts as causation in biology and physics. This process is analogical reasoning. Then (Sect. 3), I will turn to more critical issues by arguing that Faye's description of a causal factor as a difference-maker does not capture what is understood as a cause in explanations lying at the heart of physico-biological investigations, i.e. explanations in systems biology. In the last section of this paper, I will compare Faye's notion of "causal process" with the notion of "causal mechanism" ubiquitous in the biosciences. By highlighting crucial differences between the two I show why the very same cognitive criteria of causation cannot grasp both, processes and mechanisms.

2 Analogical Reasoning About Causes in Non-humans, Biology and Physics

According to Faye's explanatory strategy for identifying relevant causal factors, biologists as well as physicists conceive of causes as being *external* to a causal process under study. They initiate or trigger the causal process and bring about a change in its status. However, there is another reasoning strategy closely linked to Faye's idea that "causes make a difference and that this difference results in an effect", which should be considered as an additional cornerstone of (pre-)humans' commonplace understanding of causation and particularly of biologists' and physicists' notion of causation. This addition is *analogical reasoning* about causal relationships.

Faye notes that the development of a concept of causation in higher vertebrates like birds includes the ability to subtract features from actual actions and events. This type-level understanding of causal dependencies is based on careful comparisons of similar – actual and past – actions and events. It enables these animals to precisely plan and fine-tune their actions (e.g. to bring about similar effect in the future in different situations). By taking this relationship between pre-human causal reasoning and similarity comparisons as a starting point, the following extended characterization of a *difference-maker principle* (DM) showing potential relevance to explanatory practices in physics and biology can be developed. It takes the following form:

(DM) Event *A* is considered to be a cause of event *B* under the circumstances *C* iff, (i) event *A* would have been different, event *B* would have been different as well and (ii) one of the following conditions holds:

- event *A** (i.e. an event similar to *A*) is known to bring about a change in the event *B* under the circumstances *C*,
- event *A* is known to bring about a change in the event *B** (i.e. an event similar to *B*) under the circumstances *C*, or

– event *A* is known to bring about a change in the event *B* under the circumstances *C** (i.e. circumstances similar to *C*).

Here changes in *A* may be brought about by bodily or hypothetical manipulations or interventions. Condition (i) expresses the general idea that tracing causes includes first creating, imagining or finding a difference-to-be-explained and then to present an *explanans* that is able to account for this difference (see Woodward 2003; Waters 2007; Ylikoski and Kuorikoski 2010).[1] According to the conditions listed in (ii) something counts as a cause, if we know of one or many similar causes bringing about the phenomenon under study (i.e. producing a change in this phenomenon), if we know that a change in a similar effect as the one under study may be brought about by the event we are currently manipulating, or if we know that a putative cause event will change a putative effect event under a similar, yet different causal background.[2]

To illustrate these conditions, let us consider Faye's example of a pair of ravens hunting red crossbills. These ravens may acquire behaviorally relevant information about causal dependencies (i.e. information on how to bring about a wanted effect) if one or many of the following three cases are observed or experienced bodily:[3] a third raven (or a different raptor) interrupts the flight of a hunted red crossbill and the crossbill crashes into a plane of glass; the male raven interrupts the flight of a different bird (i.e. not a red crossbill) and this bird crashes into a plane of glass; the male raven chases a red crossbill down an alley of two different buildings (or any other obstacles) and the crossbill crashes into a plane of glass.[4] Faye's examples of causal cognition in higher vertebrates can easily be understood as cases of non-human *analogical reasoning* about potential difference-making factors as described by (DM). This strategy enables them to easily adjust their actions in order to bring about certain wanted effects in the future.

However, if we assume that (DM) is part of what Faye describes as humans' "embodied sense of causation" we may ask if it is also relevant for reasoning about causes in a more abstract way in scientific explanatory practices in the fields of

[1] In non-explanatory contexts relevant, for example, for non-human decision making, condition (i) means that an animal has to plan and/or carry out an action that is able to account for a certain desirable difference in an event.

[2] Thus condition (ii) includes similarity considerations of dependency relations showing both invariance and stability. The latter case listed above deals with similarity of stable causal dependencies.

[3] For the sake of the argument, let us assume that Faye is right about his claim that many higher vertebrates have a rich notion of causation similar to that of human beings. Recently this claim has come under attack by authors arguing that chimpanzees and other higher vertebrates are incapable of detecting underlying causal structures of events and that they merely respond to superficial perceptual cues (see, e.g., Povinelli 2000; Penn and Povinelli 2007). However, whether or not this is the case is an empirical issue and not subject to this paper.

[4] Of course, all this three cases should be understood as cases that enable observation and/or bodily experience of difference-making in order to satisfy (DM).

physics and biology. Can we thus consider the principle (DM) as another common denominator of many forms of commonplace causal reasoning and of physicists' *and* biologists' explanatory practices?

At least in biology, explanation often seems to be closely linked to similarity considerations of potential causes and causal relationships. Here are some examples: the comparison of *similar* – genetic, cellular or physiological – difference-making factors (i.e. similarity between events *A* and *A** in principle DM) is crucial for explanation and extrapolation in many research programs in systematics and taxonomy. Similarity considerations between events *B* and *B** appear in physiology in the form "A certain physiological factor *A* in animal *X* counts as a cause of a certain phenotypic feature of *X*, called *B*, if *A* is known to bring about a *similar* effect (as the one under investigation), called *B**, in animal *X** (which is, e.g., closely related to *X*)". In addition, analogical reasoning about different causal backgrounds *C* and *C** can be found in evolutionary biology, where investigations of potential causal factors of trait heritability consider artificial, yet *similar* background circumstances (e.g. in selective breeding set-ups).

Analogical reasoning about causes and similarity comparisons of different causal scenarios are also common in physics. For example, in the fluid model of electricity, changes in an electric resistor (i.e. *A* in principle DM) are understood as causes bringing about changes of quantities within a DC electric circuit *similar* to changes of resistance to flow (i.e. changes in a severe constriction; *A** in DM) in a water circuit. Drawing analogies between events *B* and *B** in order to better understand their difference-making causes is essential for the concept of self-induction: since the delayed lighting-effect of a bulb in a circuit (with inductance) is *similar* to the delay or inhibition we find in the case of a conductor in an electromagnetic field (i.e. an eddy current brake), a change in an electric current in a coil of wire counts as a cause that brings about a magnetic field which induces a voltage in the circuit itself. Here, the similarity of these effects is the systems' "inertia" in the case of change. In addition, similarity considerations of different causal backgrounds *C* and *C** are important for the Photoelectric- and Compton-effect. Here, a photon of a specific wavelength is considered to be a cause of the particle-like movement (i.e. momentum, energy, etc.) of an electron, with which it collides, since this collision has been tested with the same entities (photons, electrons) under *similar*, yet different circumstances. The difference results from the fact that in Einstein's case the electrons are bound in a sodium metal, whereas in Compton's case they are stationary and quasi free.

These examples support the idea that difference-making could be regarded as a natural partner of analogical reasoning and similarity consideration in any concept of causation claimed to play a role in higher vertebrate's action planning as well as in scientific explanatory practices. The general principle (DM) suitably summarizes this idea. However, one major problem remains. If – as Faye has emphasized in his discussion of embodied causal reasoning in animals and as I have underlined with respect to scientific causal explanations – our understanding of causal dependencies involves grasping them as similar to one another, the underlying (and unifying) concept of similarity has to be specified here. We thus have to ask how the ways

of assessing similarities between bodily experienced or observed events in higher vertebrates are comparable to claims about similarities between events in biology and physics.

In order to avoid that non-relevant, yet similar causes, effects, and circumstances are considered to trace an appropriate difference-maker and a causal dependency relation in scientific explanation, a rather strong version of similarity has to be adopted. Although developing such a concept in detail is beyond the scope of this paper, some rather brief comments may be added here: the relevant sense of similarity must be different from Lewis' (1979) discussion of the similarity of different actual and counterfactual scenarios. Lewis' account considers similarities between actual scenarios and "closest" worlds, in which we evaluate a counterfactual whose antecedent is not true of the actual world in order to gain explanatory relevant information about difference-making changes of values of a cause variable (see also Woodward 2003). By contrast, the similarity considerations discussed here should provide information about what particular *sort* or *type* the *explanantia*, *explananda*, and the causal background of a dependency relation under study belong to.

Recent approaches (see, e.g., Schurz 2009, 2011) to dealing with similarities between theories (especially in the exact sciences) by focusing on similar theoretical expressions or other structural correspondences are not suited for this issue at stake since the *relata* of causation are *events*, not theories, models, or any other concepts.[5] Thus, etiological explanation, even in highly abstract contexts, has to trace dependency relations between events and, given (DM) is correct, similarities between events. In addition, specifying the notion of similarity of causal scenarios in scientific context where no physical manipulation is possible requires detaching it from similarities in people's (and non-human higher vertebrate's) experiences of agency.[6] These and other issues may be seen as a starting point to establish a philosophical framework, which makes sense of the various forms of non-human and scientific analogical reasoning about causes as described by the general principle (DM).

3 Causes That Do Not Make a Difference

I will proceed with a critical remark on the view that causes bring about a change which biologists and physicists share independently of their specific explanatory interests. Therefore, let us take a closer look at causal explanation in systems biology – a topic only touched upon by Faye. By making extensive use of mathematical modeling techniques this field is often considered to promote a new

[5]This issue will be discussed in detail in Sect. 4.

[6]For this similarity claim of agency theories of causation, see, e.g., von Wright (1971).

"physicalization" of biology (see, e.g., Calvert and Fujimura 2011). I will argue that, although Faye's notion of a causal factor as a difference-maker is a highly suitable description of what counts as a cause in many field in the sciences, it does not capture what is understood as a cause at the modern physico-biological interface.[7]

In systems biology as well as in some fields of developmental and evolutionary biology, listing explanatory relevant information in an *explanans* often means including causes that do not make a difference or causes with very low causal relevance. Such an *explanans* is considered to address an *explanandum* only on the supposition that the former also cites those causes that, if manipulated, do not bring about a change in the latter. What is the nature of such an *explanandum*? Systems biologists often ask questions like "Under which set of causal factors does the biological system under study display stability or robustness?". This interest in the maintenance of specific functionalities of living systems under perturbation has a long history in developmental biology, where it often goes by the name of canalization (see Waddington 1942). Currently, in addition to systems biologists' approaches of dynamical modeling of equilibrium (see Gunawardena 2010; Huang 2011), the notion of robustness is most prominent in evolutionary developmental biology (evo-devo; see Hallgrímsson et al. 2002).

As recently claimed by Gross (forthcoming) with regard to systems biology, what is explanatory relevant in these cases of robustness are non-change-relating relationships.[8] To clarify this claim, let us briefly consider a typical investigation of dynamic stability in systems biology. Sudin Bhattacharya and colleagues (2011) seek to understand how gene-regulatory networks determine the stable or robust differentiation of pluripotent stem cells into unipotent cells during embryonic development. Therefore they compute the dynamics of a bistable two-gene regulatory network (i.e. a network of the two genes x and y with two attractors A and B) with mutual inhibition of the genes (see Fig. 1). They "use stochastic simulations to show that the elevation of this computed landscape correlates to the likelihood of occurrence of particular cell fates" (Bhattacharya et al. 2011, p. 2), i.e. the landscapes topography correlates to stable steady states or attractors of the network. In this model, the network dynamics of the gene-regulatory network bring about two stable pathways of development (leading to the two attractors A and B) which may be formally described.

[7]This argument includes, that also the extended version of difference-making described above (i.e. principle DM) is not able to cover all cases of scientific investigations of causation.

[8]Gross' claim primarily refers to constitutive explanation in systems biology by means of presenting a causal mechanism able to produce a system-level phenomenon which shows robustness. Some component parts of this mechanism display non-change-relating relationships or very weak dependency relations with the *explanandum* phenomenon. Below, I will understand this view as a general – causal and constitutive – explanatory strategy of citing causal events and causal capacities (i.e. properties) of entities as *explanans* variables that do not (or do very weakly) make a difference in a phenomenon under study.

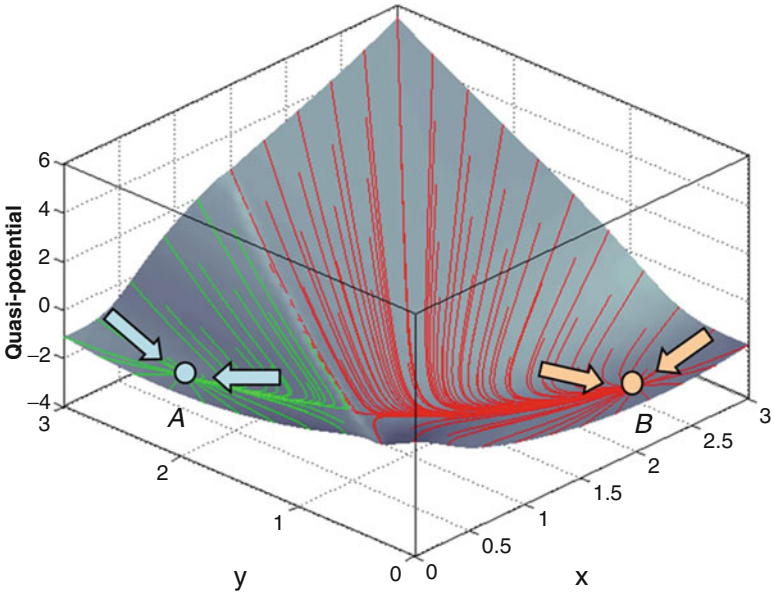

Fig. 1 Computed attractor surface of a two-gene circuit's network dynamics. Multiple trajectories converge to the two attractors *A* and *B* (*arrows*). A quasi-potential surface (i.e. the elevation represents a path-integral quasi-potential) is derived directly from deterministic rate equations describing the dynamic behavior of the two genes. (Bhattacharya et al. 2011, p. 4, original figure slightly changed; Reproduced with permission from BioMed Central)

What is considered to be an *explanatory relevant causal factor* within this mathematical model? Mainly those factors on which manipulation or perturbation – a change in the gene expression profile in one of the genes *x* or *y* at a certain point in time – does *not* change the fate of a differentiating cell. Manipulations on the gene-network are depicted in Fig. 1 as changes in the starting points of trajectories. While very few of these changes or perturbations make a difference in the *explanandum* phenomenon (i.e. they lead to a switch of the "direction" of the developmental pathway, e.g., from attractor/cell type *B* to *A*), most changes do not. Here the basic idea is that maybe all possible manipulations performed on one causal factor (e.g. on gene *x*) alone are not able to make a difference in the overall system (e.g. a change in the end point of a pathway from one cell type to another). To understand a complex system's dynamic stability, tracing such non-change-relating relationships or dependency relations of causal factors that do not bring about a change in the effect are considered as highly explanatorily relevant.

In order to specify which *explanadum* phenomenon *non-difference-makers* are thought to address in this case, we have to replace the toy model presented above by a real-life scenario. Imagine a gene-regulatory network containing 1,000 genes, which determine a bistable cell differentiation process. Now add to this picture another one million potentially perturbing (e.g. environmental or epigenetic

regulatory) factors. Now imagine that in some states at a particular point in time during cellular differentiation it is easy to switch a system from one cell type to another (*low robustness*) whereas in other situations it takes strong multifactorial perturbation to change the system's state (*high robustness*). Biologists working on robustness claim that in order to grasp the stochastic nature of a dynamic stability phenomenon one has to "embrace" the complexity of its underlying network. Models that consider only a small set of strong difference-makers – e.g. a few genes and some additional factors influencing the mapping between genes and the system behavior – often struggle to explain why the system's stability changes so radically over time under quite similar looking circumstances of causal influence. Thus, in order to solve this problem additional non-change-relating relationship and weak difference-makers are included into the model.

The strategy described here is to consider as well those "tiny" factors that may merely lead to changes on some intermediate levels. Although they do not (or only very weakly) influence the overall system's behavior under study they may mediate other stronger and (more) direct causal dependencies. Changing these weak mediating causes in isolation (one by one) does not make a difference in the overall dynamic behavior of the system.[9] However, they are crucial for understanding the system's dynamic behavior. This systemic perspective on causes may be visualized by a metaphor of an orchestra being able to play an overture, only if every musician contributes, no matter how small her part is or whether it can be heard by the amateur ear at all.

As this case of explanation of robust complex living systems shows, some scientific investigations need to trace not only information about difference-making but also dependency relations containing non-difference-makers. However, this information about which causes under which circumstances do not make a difference does not fit with Faye's notion of a causal factor as a difference-maker and his cognitive criteria of causation, respectively. Systems biologists, developmental biologists, and evo-devoists interested in explaining robustness and stability do not always understand causes in the way described by Faye.

One may thus ask: does this mean that these people act in a way that conflicts with their phylogenetically "engraved" understanding of causal factors (if there is such thing)? Maybe not. Maybe there is an evolutionary story here waiting to be told, a story that is able to make sense of our commonplace intuition that when dealing with mosaic-like phenomena of potential multifactorial causation one should not necessarily expect changing the effect by changing a single causal factor. Removing a blade of grass from a stack of hay does not make us expect the stack to change its overall shape.

[9]However, this does not mean that these causal factors are not able to make *any* difference at all. For example, changes in the expression profile in one of the 1,000 genes of the gene-regulatory network may lead to a change on the protein level (although not to a change in the dynamics of cell differentiation). But with regard to the level of the *explanandum* phenomenon these factors have to be considered as explanatory relevant non-difference-makers.

4 Causal Processes and Causal Mechanisms Compared

Faye notes that, in the sciences, the notion of a cause or a causal process in part depends on the vocabulary of the particular science and the phenomena under study. However, as described above, he also emphasizes that independently of discipline-based explanatory interests and/or the specific vocabulary used, certain dependency relations count as causal, if these relationships meet our innate criteria of causation. According to this cognitive schema, we try to grasp causes as external determining factors of causal processes and additionally, we focus on understanding the nature of *causal processes* itself. I will now address a problem linked to the notion of causal process presented here.

According to Faye, a physical causal process and a biological causal mechanism are close relatives. Both may be understood as specialized derivatives of humans' innate cognitive sense of causation. However, while some approaches of causal processes describe mechanisms as a network of interacting *processes* (see, e.g., Salmon 1984), more recent mechanistic approaches think of mechanisms as *systems* (Glennan 2002, 2009) or *structures* (Bechtel and Abrahamsen 2005) with particular causal capacities. Although, according to proponents of this new mechanism movement in philosophy of biology, mechanisms are usually conceived as *causal* mechanisms – they are said to enable the identification of causal relations – tracing mechanisms should not be conflated with causal explanation. Mechanisms are the *explananda* of *constitutive* explanations, not of causal explanations.

Let me explain this distinction: in scientific explanatory practices the notion of mechanism usually refers to causal capacities of *things*, i.e. organized systems of parts or structures (e.g. a phenotype of a cell).[10] In contrast, the notion of a causal process refers to a *series of events* spread over time (e.g. a molecule passing a membrane). This is the case because constitution is a synchronous relation between *relata* that cannot be conceived as independently existing and causation is an asynchronous relation between *relata* that can be conceived as independently existing.[11] Due to this metaphysical difference, it is hard to think of a series of events as a thing. In addition, this analysis illustrates why it is very unlikely that the *static* nature of mechanisms is being unveiled by the very same cognitive schema shaped to grasp the *dynamic* nature of causal dependencies between events, as argued by Faye.

[10]Many philosophers conceive of mechanisms as real things which can be traced in the world. On the fallacies of giving the notion of mechanism an ontic, rather than a heuristic reading, see Nicholson (2012). See also Kuorikoski (2009).

[11]On the metaphysics of constitution and causality, see Ylikoski (2013).

5 Conclusion

Taking a metadisciplinary perspective on the notion of causation becomes increasingly important as the recent debate on causation and causal explanation progressively focuses on discipline- and even sub-discipline-based phenomena in highly specific explanatory contexts. In order to not miss the forest for the trees in this debate, we have to attempt a characterization of an all-encompassing, metadisciplinary (and maybe even science-exceeding) notion of causation. I appreciate that Faye's approach pursues exactly this crucial goal. However, I am rather sceptical whether the criteria introduced by Faye suitably capture what counts as causation in biology as well as in physics.

I have argued for three points in this paper: (i) Faye's understanding of a causal factor as a difference-maker can possibly be expanded by considering similarity considerations and analogical reasoning strategies about causes in and outside the sciences. (ii) At the same time, however, Faye's description of a difference-making factor of a causal process does not capture what is understood as a cause in all fields of science. In many biological and especially biophysical investigations of robustness of living systems, causes do not necessarily have to bring about a change. (iii) In addition, a causal process should not be conflated with a causal mechanism. Elucidating the metaphysical differences between causation and constitution helps to understand why the very same cognitive schema of causation (of humans or even of all higher vertebrates) is unlikely to grasp both of them, pace Faye.

Acknowledgement I thank Dan Brooks, Jessica Pahl, Helmut Pulte, Jani Raerinne and Marcel Weber for constructive comments on earlier versions of this paper. Financial support from the Ruhr University Research School (RURS) is gratefully acknowledged.

References

Bechtel, W., and A. Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Science, Part C* 36: 421–441.

Bhattacharya, S., Q. Zhang, and M.E. Andersen. 2011. A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC Systems Biology* 5: 1–11.

Calvert, J., and J.H. Fujimura. 2011. Calculating life? Duelling discourses in interdisciplinary systems biology. *Studies in History and Philosophy of Science, Part C* 42: 155–163.

Glennan, S. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69: S342–S353.

Glennan, S. 2009. Mechanisms. In *The Oxford handbook of causation*, ed. H. Beebee, C. Hitchcock, and P. Menzies, 315–325. Oxford: Oxford University Press.

Gross, F. forthcoming. The relevance of irrelevance: Explanation in systems biology. In *Explanation in biology: An enquiry into the diversity of explanatory patterns in the life sciences*, ed. C. Malaterre and P.-A. Braillard. Berlin: Springer.

Gunawardena, J. 2010. Models in systems biology: The parameter problem and the meanings of robustness. In *Elements of computational systems biology*, ed. H.M. Lodhi and S.H. Muggleton, 21–47. Hoboken: Wiley.

Hallgrímsson, B., K. Willmore, and B.K. Hall. 2002. Canalization, developmental stability, and morphological integration in primate limbs. *American Journal of Physical Anthropology* 35(Supplement): 131–158.

Helmholtz, H.v. 1896. Über das Ziel und die Fortschritte der Naturwissenschaft. In *Vorträge und Reden*, vol. I, 5th ed, ed. H.v. Helmholtz, 367–399. Braunschweig: Vieweg.

Huang, S. 2011. Systems biology of stem cells: Three useful perspectives to help overcome the paradigm of linear pathways. *Philosophical Transactions of the Royal Society of London, Series B* 366: 2247–2259.

Kuorikoski, J. 2009. Two concepts of mechanism: Componential causal system and abstract form of interaction. *International Studies in the Philosophy of Science* 23: 143–160.

Lewis, D. 1979. Counterfactual dependence and time's arrow. *Noûs* 13: 455–476.

Mach, E. 1897. *Contributions to the analysis of sensations*. Trans. C.M. Williams. Chicago: Open Court.

Mach, E. 1910. *Popular scientific lectures*, 4th ed. Trans. T.J. McCormack. Chicago: Open Court.

Nicholson, D. 2012. The concept of mechanism in biology. *Studies in History and Philosophy of Science, Part C* 43: 152–163.

Penn, D., and D.J. Povinelli. 2007. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology* 58: 97–118.

Povinelli, D.J. 2000. *Folk physics for apes: The Chimpanzee's theory of how the world works*. Oxford: Oxford University Press.

Salmon, W.C. 1984. *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Schurz, G. 2009. When empirical success implies theoretical reference: A structural correspondence theorem. *The British Journal for the Philosophy of Science* 60: 101–133.

Schurz, G. 2011. Structural correspondence between theories and convergence to truth. *Synthese* 179: 307–320.

Vollmer, G. 1985. *Was können wir wissen?* Die Natur der Erkenntnis, vol. I. Stuttgart: Hirzel.

von Wright, G. 1971. *Explanation and understanding*. Ithaca: Cornell University Press.

Waddington, C.H. 1942. Canalization of development and the inheritance of acquired characters. *Nature* 150: 563–565.

Waters, C.K. 2007. Causes that make a difference. *Journal of Philosophy* 104: 551–579.

Woodward, J. 2003. *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Ylikoski, P. 2013. Causal and constitutive explanation compared. *Erkenntnis* 78(2): 277–297.

Ylikoski, P., and J. Kuorikoski. 2010. Dissecting explanatory power. *Philosophical Studies* 148: 201–219.

Just Complexity

Max Urchs

1 Meanings

Complexity is admittedly hard to explain. It seems strongly connected with all sorts of intriguing concepts such as deterministic chaos, surprise, non-linearity, dialectical leaps, phase transition, surprise or unpredictability, distributed and locally generated order, organization and control among numerous elements. A system can already be regarded as complex when it allows for emerging properties. For instance, a nervous system is complex if it can generate consciousness. Such a characteristic presupposes the existence of emergent properties. Since emergence is knowledge-dependent, so too is complexity. Clearly, the aforementioned properties are of remarkably inhomogeneous nature. Some, such as non-linearity, can be formally defined; others cannot, at least at the current time, though a precise description is indeed available (see e.g. recent work on chaos). Surprise, to take the other extreme, seems hardly capable of being expressed in formal terms at all. All this suggests that the underlying intuitions are very diverse indeed. There are many ways by which the complexity of a system can be estimated. The mere synopsis and classification of complexity measures is a Herculean task in itself. In order to gain an orientation, Melanie Mitchell's "Guided Tour" to complexity (Mitchell 2009b) is still a very rewarding read. Another position which can be recommended for an overview is that of James Ladyman et al. (2011), where he provides a concise definition:

> A complex system is an ensemble of many elements which are interacting in a disordered way, resulting in robust organization and memory. (Ladyman et al. 2011, p. 27)

That, however, is a courageous exception. In general, after having diagnosed the multiplicity of intuitions and meanings, authors proceed to an overview of

M. Urchs (✉)
Department of Strategy, Organization & Leadership, EBS Universität für Wirtschaft und Recht, Gustav-Stresemann-Ring 3, 65189 Wiesbaden, Germany
e-mail: max.urchs@ebs.edu

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_14,

possible options for a definition of complexity. They are to be found, naturally, in the fields where complexity has been an issue for years. These include, on the one hand, statistical physics and, on the other, algorithmic complexity theory. The first deals with subjects such as the rotating waves of heated fluids while the second focuses on e.g. cellular automata. Nevertheless, it may still be felt that these do not reflect our intuition correctly. For lack of better terms, let us designate the kind of complexity that we encounter in the external world as *real complexity*, and the abstract counterpart thereof, which is processed in formal frameworks, as *plain complexity*.[1] Phenomena such as heat waves and cellular automata are rather metamathematical counterparts of what we intuitively take to be complexity; they represent cases of plain complexity. Real complexity refers rather to various second-order dynamics: the underlying structure of the phenomenon is in flux. The very mechanism that generates a system's behaviour is changing, leading thereby to the evolution of new forms of connections. The prototype model for complexity should certainly not be the chaotic pendulum.

What has been said so far certainly does not amount to a succinct and satisfactory definition of complexity. It may not be possible to give one. Complexity can perhaps be understood only in the actual experiencing of it. This is common to many concepts, e.g. intelligence, creativity, consciousness, etc.

If no adequate definition of complexity can be provided, it is probably best to attempt to illustrate it in terms of a metaphor. It will not be an example from the sciences, since we think that it is more advantageous not to confine the intuition to one particular scientific discipline. Think about playing jazz. There are some styles of music, for example classical and folk music, where the notes played are set and invariable. In other words: they are always played in the same way with no variation of form. Jazz music, on the other hand, is an improvised music. Starting out with a set form – a popular song or dance tune – the performers will then improvise over the form. The improvisations can be seen as an open-ended series of possible variations. External influence – the audience, the setting, the musicians' mood – will all also effect the performance, which can never be identically reproduced. If you were to ask a practicing jazz musician to define the complexity of the improvisational process, he would be at loss to do this adequately. He could certainly be able to give us much formal information about harmonic theory and the like, but then, if pressed for a definition of how decisions are made in the actual improvisational process, the most he could reply, with a shrug of his shoulders, would merely be: "I just do it!"

Within the last decade – owing predominantly to the work of Sandra D. Mitchell (2008, 2009a) – a further meaning of the word has emerged: complexity as denoting a way of looking at reality, as a *Weltanschauung*. Complexity as *Weltanschauung* implies a robust position concerning the essence of scientific model building. It assumes that mental life, social reality and much of physical nature are ineluctably

[1]The connotation of "plain", i.e. "simple" is intentional: Complexity in models – though mathematically sophisticated – is simpler than in reality. *Simplexity* would be a better name but it is already in use.

complex. According to such an opinion, complexity is a fundamental and ubiquitous property of reality. We will use the term complexism[2] to refer to that very general position.

2 Causes

To come back to the nature of complex systems, there is something problematic about plain complexity. The problem is twofold. Firstly, plain complexity as characterized above excludes causal explanation. However, explaining things in causal terms is crucial for a common understanding of science.

Let us begin with the latter point. In complex environments we hardly ever encounter the traditional ball-on-the-pillow causality. Quite often it is some rather fuzzy form of causal propagation, massively conditioned and full of exceptions. Instead of strong causality in a Humean sense we encounter a plurality of non-homogeneous causal threads, which interlace somehow to a causal braid connecting cause(s) and effect(s). The principle of strong causality breaks down. The metaphor that determines causal intuition is more often the notorious one of the butterfly in the Amazon destroying budding landscapes in Texas rather than the classical one of billiard balls.

According to some authors (e.g. Lipton 2005), not much hinges on the use of a causal *façon de parler*. They therefore recommend avoiding causal terminology altogether and, instead, reporting the results of empirical research in a non-causal way. In my opinion, a strategy that completely rejects causal phrasing is not sensible. A causal understanding of the environment is vital for human beings. It may even be specific to us and it therefore plays a role in all sorts of cultural activity, including science. Not in all branches of science, to be sure. There are well-known cases in which causality is not taken into account by a specific scientific discipline. Nevertheless, and although causality is irrelevant in some areas, Russell's notorious judgment should certainly not be generalized for all of science. We also see a renaissance of causal thought in physics.[3] In general, one should keep searching for causal nexus and causal laws, since that is what science is supposed to do.

Whatever the mathematical details of defining (plain) complexity are, it is important to note that the complex behaviour of a system can be precisely stated only after appropriate mathematical modelling of the system has been carried out. To say that some real-world system displays complex behaviour thus presupposes that we have its description at a reasonably high level of abstraction. Causality does not play any role in mathematics. Functional dependency is all which remains in highly abstract areas of scientific thought. Not until such a level of abstraction has

[2]Also "complexism" is already in use (as is "complexicism"). However, it is used in a somewhat strange way. I would be inclined to redefine the notion for the present purposes.

[3]The entry "causality" is by far the largest in the subject index of Esfeld's (2012).

been achieved will complexity really come into its own, and causality will, under these circumstances, become redundant.[4]

All this indicates that the above attempt to define plain complexity barely begins to exhaust the essential intuition concerning real complexity. In particular, complex behaviour should include causal aspects. It thus amounts to a double characteristic: a real complex system contains causal elements and is transformed into a plain complex system by further abstraction. So what are these causal elements?

In her book (Mitchell 2009a, b) Sandra Mitchell discusses in detail the case of a major depressive episode in order to provide an understanding of the specific form of causality in complex circumstances.[5] Mitchell elaborates her view on a causal nexus in a complex environment partly as a commentary on a paper by Dominic Murphy (2008). This work provides an illustration of causal pluralism in psychiatry which may well be generalized to other disciplines. According to Mitchell, a reduction – as entailing a "nothing more than" account of what is causally significant or sufficient – fails to capture important aspects of complex systems. This does not mean that reductionism should be dismissed altogether.

> Rather than defining a comprehensive antireductionism as the antidote to the seemingly obligatory reductionism, [we need] a pluralistic view of valuable explanatory strategies employed by contemporary science, which includes reduction. (Mitchell 2008, p. 22)

The general framework that seems appropriate for the patchwork of laws in a dappled world is a pluralistic approach towards causality. There are various forms of causality in the sciences and there are many cases in which it is necessary to apply more than one form to arrive at an acceptable model. In "Causal Pluralism", Stathis Psillos convincingly argues against what he calls the straightjacket, namely that

> a good philosophical theory of causation should tell a unified and complete story that covers each and every aspect of the nature of causation. (Psillos 2010, p. 133)

Nowadays there are numerous publications about the general demand of causal plurality and about the technical details of an appropriate framework. Phil Dowe's process theory of causation, e.g. (Dowe 2000), Huw Price's causal perspectivalism, or: standpoint causality (Price 2001, 2007), Chris Hitchcock's causal scenarios (Hitchcock 2003), Leen De Vriese's epimethodological causal pluralism (De Vreese 2006), Johannes Persson's work on polygenetic effects (Persson 2007), or my own conception of causality based on episystems (Urchs 2001) are all examples of frameworks for multiple causal nexus. The main problem is, of course, to find out how to integrate all these partial and varying causal connections into one

[4] Of course, if the abstraction is sufficiently weak, then we end up without any sensible notion of causality as well. Everything that appears in such a world-view is merely chaos, a wild mess. In the "Ursuppe" (primeval soup) of universal interaction it is impossible to discern any properties or relations. No individual causal dependencies can be isolated.

[5] Instructive as this is, it did not prevent the authors of the recent edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) from searching for a classification based purely on biological factors concerning the genetics of psychic disorders. Ironically, David Kupfer, the coordinator of DSM-revision, is fellow faculty with Sandra Mitchell.

sufficient causal story. In general, that story will not be consistent. But modern logic can cope with such a situation. More then half a century ago, it came up with formal calculi which tolerate inconsistent sets of assumptions. Nowadays, there is an increasing number of inference engines which lead to reasonable consequences when applied to inconsistent premises. Of course, one should always make things in science as consistent as possible at a given stage of theory development. But if any inconsistencies are subsequently found to be present then one should accept them. It helps to know that inconsistencies can be tamed and controlled.

A similar change in attitude should occur with respect to real complexity in the sciences. Elaborated formal methods for handling plain complexity give grounds for confidence in the feasibility of researching real complexity in a like manner, as did paraconsistent logic with respect to inconsistency.

3 Markets

Let us now look at a first group of examples of how complexity takes effect in the sciences. In recent years, there have been quite a number of approaches in the economic sciences which address seriously the issue of complexity. Many of them, especially in finance, originated in one way or another in the work of Benoit Mandelbrot. The central idea is to apply to economic phenomena various models and concepts associated with the physics of complex systems – e.g. statistical mechanics, condensed matter theory, self-organized criticality, microsimulation, etc. A second approach takes biology rather then physics as its starting point. "Econobiology" ("evolutionary economics") attempts to employ the insights of evolutionary biology in order to explain economic phenomena. Economic complexity is considered as analogous to biological complexity. However, all methods should be carefully adjusted to the new area of application. To suppose that a method which has proven to be of value in some one area will continue to function just as well in another means deriving functional equipollency from the similarity of outward appearances. In other words, it means magical thinking.

Even sophisticated adjustment may not be enough to yield adequate models. Economic theory is a primary example. This pursuit of formal elegance can be observed in the very beginnings of economic science. When creating "social physics" (as economics as a scientific discipline was called), Adam Smith – being, after all, a moral philosopher – was perfectly aware that *homo oeconomicus* does but poorly reflect human economic agents. In this way, the heroic entrepreneur, the proletarian struggling to survive from day to day, and the greedy banker, have all been demoted to the level of autistic utility maximizers. It was clear to him and to most of his followers that such a model of economics would hardly be able to grasp the essential features of economic reality. But *homo oeconomicus* set the parameters for quantitative modelling. And, in academia, the use of mathematical methods was crucial for acknowledgement as a mature science. At that time, physics was overwhelmingly successful. So it was only natural for emerging disciplines,

such as psychology or economics, to emulate that success story by copying the method of physics, that is, the processing of hard facts by quantitative procedures. In psychology, this strategy led mostly to nonsensical results. Economics was somewhat better off since the basic facts of production and of the distribution of economic goods are of a quantitative nature. But even if nineteenth century economists had used the most advanced mathematical methods available (which they never did), attempting to embrace all the violence and passion of early capitalism in these simple models would have been a futile endeavour. The pressure to use quantitative methods will certainly not have been the only reason for this failure. Actually, there seems to be an inclination towards simplicity, as simplicity, in scientific contexts, is associated with elegance, with aesthetic appeal. According to some authors, beauty of that kind is inherently related to (pragmatic) truth – we like what works. Consequently, pleasing things are the right things. That might be another reason why scientific work tends to evolve towards abstract beauty, even at the peril of drifting away from applicability. That pursuit of elegance occasionally leads us astray.

Economic models based on simple assumptions may simulate the evolution of some specific economic indicators to an amazing degree. But they are purposefully trimmed to do so. Under slightly different conditions, which may cause a dramatic shift in the real system's behaviour, any similarity may break down. It turns out that such formal constructions are not adequate models for real systems – their "range of forecast" is all too narrow. This is exactly what happened with economic models in the current crisis. If central features of the system have no essential counterparts in the model, then the model will work only under very specific conditions. It is like a compass that will work properly only if there is no wind. Navigational instruments, however, are especially needed when the weather becomes stormy. The analogy with the instruments for economic forecast suggests itself.[6] In other scientific disciplines, the same is true accordingly, though perhaps with less significant consequences. Transplanting methods from statistical physics to finance may enhance a feeling of omniscience and almightiness in some minds. In his 2004 speech "The Great Moderation", Ben Bernanke declared the age of volatility and incomputability in financial markets to be over. Some years later, the Wall Street experts accidentally caused the largest financial collapse ever seen, which has led to the worst global economic crisis since the Great Depression.

For the pricing of a stock-market portfolio, neoclassical economic theory used to assume an analogy between the decisions of economic agents and stochastically moving independent particles. It was soon observed that this was inadequate. The model does not capture extreme movements of the market. Unfortunately, Black Swan Events like the crash in 1987 – unforeseen, with extreme impact, retrospectively predictable – did happen far more frequently than the model predicted. Benoit Mandelbrot lived to see his enduring criticism of the models justified. Economic agents, unlike molecules, do react to each other and may crowd into herd

[6]Therefore, Milton Friedman's opinion, not to be concerned with the adequacy of models as long as they yield acceptable prognoses, is all but self-evident.

behaviour. Anticipating market dynamics may massively influence their decision making. The crisis of 2008, after Lehman Brothers went bankrupt, demolished an even more fundamental assumption. Since large financial institutions may fail, the risk of such an event has to be priced into the transaction. But no algorithms were available for calculating the risk. Inter-bank loan broke down, with dangerous implications for the real economy. Moreover, the no-arbitrage rule seemed to have been abrogated. This would mean that financial products no longer have any definite price. Accordingly, the financial crisis exposed that standard models of economic finance were too simple and had to be augmented. Such an extension, though troublesome, is possible without having to replace the basic framework of the model. The transition from traditional theory in finance to e.g. behavioural models needs an expanded preference function, but leaves the procedure of utility maximizing intact. But what if there are no stable preference orders after all?

John Coates thus describes the situation in Wall Street at the end of 2008:

> By mid-December, the financial industry has endured a month and a half of endless volatility and non-stop losses. [. . .] Many firms, facing bankruptcy, have closed their doors. One by one, the lights are going out all across the financial world. [. . .]
>
> Economists assume economic agents act rationally, and thus respond to price signals such as interest rates, the price of money. In the event of a market crash, so the thinking goes, central banks need only lower interest rates to stimulate the buying of risky assets, which now offer relatively more attractive returns compared to the low interest rates on Treasury bonds. But central banks have met with very limited success in arresting the downward momentum of a collapsing market. One possible reason for this failure could be that the chronically high levels of cortisol among the banking community have powerful cognitive effects. Steroids at levels commonly seen among highly stressed individuals may make traders irrationally risk-averse and even price insensitive. Compared to the Gothic fears now vexing traders to nightmare, lowering interest rates by 1 or 2 per cent has a trivial impact. (Coates 2012, pp. 212–213)

Basing financial policy on inadequate models may result in mistaken and very expensive decisions. I do not think it is fair of Coates to suggest that testosterone, the suspicious "molecule for irrational exuberance" is to blame for the financial crisis. Appeasing the market by lowering the level of this hormone in financial institutions seems to be too creative an idea. Nevertheless, the phenomenon described by Coates demands fundamental changes in the considered models. Individual traders, whether afflicted with high testosterone or not, do not appear in conventional models. Changing the models accordingly renders them heterogeneous. There will be various levels of description, with interactions between them. From the received perspective, learned helplessness among investment-bankers might well be considered to be an emergent property. In other words, the models will turn complex.

Among the early advocates of complexity in social theory, John Maynard Keynes plays a prominent role. Many of his unorthodox views on model building in economics show great adjacence to the ideas of recent authors on complexity.

> [. . .] we can attribute a definite measure to our future expectations and can claim practical certainty for the results of predictions which lie within relatively narrow limits. Coolly considered, this is a preposterous claim, which could have been universally rejected long ago, if those who made it had not so successfully concealed themselves from the eyes of common sense in a maze of mathematics. (Keynes 1921, p. 424)

While reading Keynes' deliberations on the essence of vagueness and particularly on the role it has to play in scientific methodology, one feels inclined to read complexity into his intended understanding of the concept of vagueness. Of course, the notion of complexity was uncommon in methodological discourse at that time. And yet, it seems that it is indeed complexity which Keynes had in mind when talking about vagueness.

> Yet there might well be quite different laws for wholes of different degrees of complexity, and laws of connection between complexes which could not be stated in terms of laws connecting individual parts. In this case natural law would be organic and not, as it is generally supposed, atomic. (Keynes 1921, p. 227)

His pioneering attitude regarding this topic was no coincidence. In the early 1930s, we see a close and fruitful intellectual exchange between the most brilliant minds of the Cambridge faculty of economics and philosophy: the lively contacts between Keynes, Piero Sraffa, Frank Plumpton Ramsey, George Edward Moore and Ludwig Wittgenstein are all well documented (see Coates 1996).

Such views were fundamental to the methodology he used in his seminal work *General Theory of Employment, Interest, and Money*.

> I argue that important mistakes have been made through extending to the system as a whole conclusions which have been correctly arrived at in respect of a part of it taken in isolation. (Keynes 1936, p. xxxii)

In the social sciences there is usually dialectical interplay between dynamics and structure. No single one of them without the other will be sufficient. Both self-regulating markets and algorithmic central planning are myths. Free markets need framework regulation. Similarly with law and arbitrariness: neither Prussian drilling nor anarchy can hold a society together. Only a mixture thereof will do. The mixture *per se* and the road to establishing it are complex phenomena. Any methodology of real complexity needs to make room for that dialectics, i.e. it has to be inconsistency-tolerant.

4 Brains

The human brain is complex in every sense of the word. Unsurprisingly, the investigation of its properties and functions has been a continuous search for a reduction of complexity. And to where has it led? In early neurosciences, the paradigm model envisioned the brain as a kind of department store. A specific brain area was allocated to each mental function or capacity, resulting in map-like segmentations of the whole brain. Over time, brain maps became incredibly precise with respect to both function and localization. One may wonder, nevertheless, about their methodological grounding. Bizarre cases of brain injury such as Phineas Gage's, not to speak of the wide practical experience of neurosurgeons, which are inconsistent with the doctrine, might have raised doubts about such a sharp functional segregation of the brain. However, they actually did little to undermine

trust in the allotted organization of the neocortex. Little by little, a clear-cut catenation between brain pathology and psychological dysfunction has appeared. The scenes on the mental stage were thus embedded into the histological topography of Brodmann's areas – the contemporary level of understanding of the physiological organization of the brain.

But it was not quite that simple. Technological progress into investigation of the brain has revealed an increasingly multifaceted heterarchy of the brain. Santiago Ramón y Cajal's indirect reply to Camillo Golgi's speech at the Nobel event of 1906 demonstrated a tendency towards greater complexity of the nervous system. That seems to indicate a general trend in brain research: instead of revealing homogeneous structures, things tend to become more complex at each new level of investigation. In 2005, Olaf Sporns (Indiana University) and Patric Hagmann (Lausanne University Hospital) independently suggested the term "connectome" to refer to the entirety of the neural connections within the brain.

> It is clear that, like the genome, which is much more than just a juxtaposition of genes, the set of all neuronal connections in the brain is much more than the sum of their individual components. The genome is an entity it-self, as it is from the subtle gene interaction that [life] emerges. In a similar manner, one could consider the brain connectome, set of all neuronal connections, as one single entity, thus emphasizing the fact that the huge brain neuronal communication capacity and computational power critically relies on this subtle and incredibly complex connectivity architecture. (Hagmann 2005, p. 123)

Because of differing levels of spatial resolution in brain imaging, brain networks are defined at different levels of scale. Objects at each level are investigated using different techniques, ranging from electron- or light-microscopy to diffusion MRI and fMRI. The ultimate goal is to integrate connectomic maps of different scales into one single hierarchical map of the neural organization of a given species. For the human brain, there are huge technical problems to be overcome. They mainly result from three sources: the sheer amount of data (given today's technologies, data collection would take years); adequate machine vision tools to annotate the data are not yet available; and the same applies to high-performance algorithms for the analysis of the resulting brain graph.

That does not mean that the problem is hopeless. Due to innovatory observational technologies (cutting nervous tissue with a diamond knife with an edge of about 12 carbon atoms wide into ultra-thin slices and microscoping the contacts of singular neurons) there has been remarkable progress in specifying (tiny parts of) the local connectome and in unravelling the structure of the giant knot of intertwined neurons into graphs representing unscrambled synaptic chains. One may justifiably hope, for example, that such a representation of a tiny region (just a fraction of a cubic millimetre in volume) in the dorsal ventricular ridge of a male zebra finch will allow scientists to establish the timeline of the spikes of its neurons and thus explain the characteristic song of this particular finch and render it technically reproducible.

In every case, a major methodological challenge for macro-scale connectomics is finding adequate parcellations of the brain. Just to partition it into equally sized regions or into anatomical regions with an unclear relationship to the underlying functional organization of the brain will not give optimal results. To make matters

worse, the functional organization of the brain – the amount of communication between different regions, also known as the functional connectome – is to some degree independent of the structural organization and changes over a much shorter time scale. In any case, the human connectome, with its interacting processes going on at different levels of the structure, provides yet another interesting case of a complex system.

Even the simplest well-investigated nervous system, the neuronal structure of *C. elegans*, displays a remarkable structural variety. Although it consists of only about 300 neurons, these neurons represent more than 100 different cell types. The worm *C. elegans* was the first organism whose connectome came to be completely known.[7] Because of the huge number of different neuron types, however, the connectome alone is insufficient to provide a complete understanding of the worm's neural processes. Knowing the organization of the cells of *C. elegans*' nervous system is as indispensable as knowing their individual workings. The functionality of a system depends on the properties of its elements and on the specific organization of the elements. The more diverging are the elements of some entity, the more the elements' properties matter in understanding the entity's performance; or, the larger the number of different types of a structure's parts, the bigger their relative impact on understanding its functionality. And vice versa – the smaller the number of different types of parts forming some structure, the higher is the relative importance of the organization of the parts in understanding the structure. Just to take Sebastian Seung's illustration: when you play with old-fashioned Lego sets (containing only one type of block), then describing your creation means enumerating the organization of its parts (see Seung 2012, p. 268).

How is the relative importance of types and organization distributed in the case of the human brain? Neocortical structures are formed by billions of neurons (and ten times as many glia cells). Every neuron has up to 10,000 synapses, each of them with millions of ion channels to run intercellular communication. In the human brain there are far less neuron types than in *C. elegans* – we know of only about 100 different types of neocortical neurons. That is why structure, the connectome, is of central interest. Knowing the connectome is necessary for understanding the brain. Consequently, there can be no correct simulation of a human brain based on a wrong connectivity. (To be sure, the part list of the human brain is relatively short, though it is certainly long enough to provide neuroscientists with years of intensive investigations.)

Although the functional modularity of brain organization used to be one of the best established tenets in the neurosciences, the kind of modularity which was assumed for the human brain has changed over time. Coming back to the analogy of the department store of the phrenologist, modern neuroscience assumes more flexibility and more interaction between the departments. Structural units alone are not very informative, because they may serve various functions, depending

[7]This means that every synaptic connection of every neuron in *C. elegans*' nervous system has been identified.

on their specific connections. This does not necessitate the revival of holism. However, light is shed on functional networks, thereby revealing a much greater organization of locally and physiologically distributed areas involved in serving a specific task. It resembles the situation in genetics: genes encode proteins which can do a lot of different things. The modularity of the human brain is certainly not any "cut it out-and-analyse its function" kind of modularity: modules are connected, rather than independent. And it is that latter kind of modularity which would obviate complexity. So it seems that traditional modularity, i.e. separatedness, loses importance in neurophysiology, as did linearity in physics. It often seems that progress in the sciences encourages complexity, rather than not.

Is there a remedy for the trend of increasing complexity in the realm of brain research? The ultimate way to reduce the complexity of an object of investigation is to build a faithful model. In the case of the human brain, attempts are as old as science itself. At every stage of technological progress, the most advanced machines were believed to have finally achieved the goal. Mills, pumps and outlets, clockworks, telephone switchboards, electrical devices, and electronic calculators were efforts to simulate the human brain, to make the machine think. Over recent decades, there have been many interesting projects in artificial intelligence. More than 10 years prior to the DARPA SyNAPSE project, Hugo de Garis directed the Robo-Koneko ("artificial kitten") project. Its objective was to create a neural network composed of tens of millions of neurons. This resource would power in real-time an external robot which was supposed to behave like a natural kitten. Many criticized the Robo-Koneko project for being too simplistic, pointing out that simply magnifying neural nets was not the solution to anything. Hugo De Garis seemed to have little interest in finding out how a cat's brain worked. Somewhat similar to an ancient alchemist, without fully understanding the underlying mechanism, he trusted in methods of evolutionary engineering which would eventually lead to an "opaque", but working machine. Though certainly very interesting and well ahead of its time, the project failed. The main reason was, I suppose, its insufficient complexity. Just putting together more and more of the same and hoping for a dialectical leap requires a lot of faith.

One of the promising recent projects does not repeat that mistake. Henry Markram is a widely renowned neuroscientist for his pioneering experiments on synapses. He is aware of how important it is to understand how a brain works in order to simulate it on a computer. In 2009, he announced a computer simulation of the human brain within 10 years. But what is the criterion for judging success? In 2019 we may use the good old Turing Test to see whether we have reached the final destination and created a thinking machine. But how do we know that we are moving in the right direction until that day comes? Markram's Blue Brain is composed of units that adequately model electrical and chemical signals in many types of neocortical neurons. But – as we discussed previously – it is connection that matters and we will not have any cortical connectome soon. At least not on the level of precision that Markram sets out to model. So there is no antetype for connecting the model neurons with each other. For the time being, researchers connect them randomly. In the living brain, however, there seems to be activity-dependent synapse

elimination and creation, which leads to a nonrandom pattern of the surviving connections. Another question concerns the subtleties of neurotransmitters in the synaptic cleft. Can neurons interact outside the confines of the synaptic cleft? Which role do hormones play? There seems to be plenty of room below the level of the connectome. Furthermore, glia cells, by definition, do not form any part of the connectome. But, as is known, they influence cognitive processes in the brain. So, although it will be very challenging to meet the required level of knowledge of the connectome, it may be that even the connectome is not enough.

From yet another perspective, one may ask whether an isolated brain is the right object to develop higher mental phenomena which are typical for humans, and particularly whether the machine will possess consciousness. In a word, it seems that for the time being the project, although an enormously complex enterprise, is still not complex enough. A possible analogy to the Human Genome Project suggests itself. Sequencing the genome, we did not find exactly what we were hoping for. It turned out that much more has to be taken into account than just the order of fourfold bases on the DNA. In that sense, HUGO opened a gateway to a much more complex world below DNA level, thus creating new branches of science.

Recently, the Human Brain Project, led by Markram, was awarded one of the two Flagship Projects in Science in the European Union and received giant funding over a 10-year period. That is a very satisfying decision by the European Commission, indeed. Not only because in the course of that mammoth project many important scientific results will doubtlessly be achieved, but also because the project will endure until it finally reveals its promised result.

5 Revolutions

What future role will complexity play in the sciences? Recently, there seems to be widespread dissatisfaction and an awareness of fundamental methodological problems in many areas of natural and social sciences. Complexity is one of the topics which provides hope for some sort of methodological breakthrough. In particular, it is plain complexity which has the potential to bridge the innovation gap – starting out with traditional models and processing them with high-powered computing machinery makes the resulting new models more familiar. For instance, the concept of *homo oeconomicus* is no longer considered adequate as an economic decision-maker. A whole swarm of these autistic zombies, however, interacting under simple rules may form interesting and relevant raw material for the massive modelling of economic decision-making. It may be somewhat adventurous to predict that complexity will develop into a central topic of the philosophy of science. And yet it focuses on many important issues such as causality, structure, predictability and order. Thinking about complexity may improve one's mental models by rendering the semantic web less complex. May it perhaps serve as a unifying framework to address general problems of the status of scientific knowledge?

What about complexity as a *Weltanschauung* – will it change our image of science? It is always difficult to assess historical trends in which one is immediately involved. First of all, participating in a process dilates its time scale. Even a process retrospectively identified as a revolution will look to those involved like a crawling revolution at the most.

In order to gain a perspective let us look at the Renaissance swerve. Six hundred years ago, conditions were difficult for science; it was undergoing a life-threatening crisis. The crisis was caused mainly by the unwillingness to engage in applied science, to contribute to technological innovations which were badly needed to make Joe Blogg's everyday life easier. There were truly extraordinary findings in medieval logic and ontology – but no conscious and deliberate application of these breakthroughs which would alleviate the aforementioned Joe Blogg's quotidian existence. There were some truly extraordinary findings in medieval logic and ontology, but there was, at the same time, an almost complete breakdown of communication between science and society. While scholastic scientists were actually pondering the psycho-physical problem, people in the towns were disgruntled by what they saw as fatuous debates in the ivory tower about angels dancing on a pinhead. Science had to regain social acceptance by moving towards applied science. Such a redirection required more than just another scientific method. Induction came along with a new criterion for scientific evidence. The rules of the game changed to a wide extent. Galilei dropping metal balls down the Tower of Pisa, though historically inaccurate, presents a superb icon for the ongoing process of change. His fellow scientists were upset. They refused to even take note of Galilei's experiments. They might have protested, "That is not how we do science! Every student knows the proper method – read the relevant works of the great precursors, debate the issue among your learned colleagues, and then find the received knowledge confirmed, or, in the unlikely event that they had erred, correct the error in an addendum. Throwing objects to see them falling down is not what a scientist should engage in." Actually, a new mode of thought was needed to bring about Renaissance science. The new science made a pact with society:

> We, the scientists, will work hard to uncover the most fundamental structures of the world. We will describe them in mathematical language in a Golden Book of Nature. Everyman receives a precise and perspicuous picture of reality that lays the foundation for technological progress. We will be rewarded for that.

Call that the *Renaissance promise*: nature is founded on a mathematical formula. We will dig it out for you.

Five hundred years later, at the end of nineteenth century, the situation was different in many respects. Leading opinion-makers announced the impending end of science. Science, in particular Newtonian physics, achieved unprecedented success. The Golden Book of Nature was nearly complete. Everything that could be discovered by science had been discovered. A few blank pages remaining at the end of nineteenth century would easily be filled in by a few talented PhD candidates. Afterwards, we would ultimately close the Golden Book and end the project "science" and devote our time to other joyful activities. We all know what

happened next. The blank pages "black body radiation" and "Mercury perihelion" were filled out in a way that turned the Golden Book into a museum piece. The eve of the final triumph turned out to be an eve of destruction. Both physics and mathematics underwent foundational crises at that time.

However, it took some decades for the detailed implications of the new discoveries in all the relevant fields to emerge, that is, in the relativistic universe, in the quantum world, and in foundations of mathematics. Gödel's and Turing's results on the range of algorithmic methods and Poincaré's and others' work stimulated by the three-body problem made mathematics a more confusing area. Even today, there is still little understanding of the standard model of modern cosmology, and not much has been achieved in the area of reconciling quantum mechanics with common sense.

It is hard to predict where the present phase of the evolution of science will lead to. The characteristic new facet seems to be (plain) complexity. Einstein's celebrated admonition not to overemphasize simplicity[8] anticipates the central idea. Nowadays, there are voices from many fields contending that the striving for simplicity has gone too far. Oversimplified elegant models in economics do not take certain essential factors into account and fail as soon as the weather turns stormy. Structural models in the neurosciences are too coarse-grained and thus fall victim to voodoo correlations. There is increasing discontent together with rising awareness of the problem not only in econometrics and neurophysiology, but also e.g. in epidemiology, anthropology, cosmology, psychiatry, pharmaceutics and oncology. The world is complex, as Sandra Mitchell puts it, and we ignore complexity in scientific theories at the peril of irrelevancy.

The new mode of thought in the Renaissance period was characterized by respect for the factual, by the concern for living practice, i.e. by scientific analysis of our environment. The new mode of scientific thought was made possible by the enhanced arsenal of scientific methods. The Renaissance revolution had established the knowledge of practice as a rewarding challenge for science. The ongoing revolution centres on the imperfectness of that knowledge. In some cases, the complex nature of a phenomenon imposes upon us the renunciation of knowledge that guarantees reliable prognosis and control. That is an imposition, indeed. We should accept reality as cognizable by human standards, as opposed to "demonic" standards à la Laplace. In many cases, there is no arbitrarily precise future prospect available. The facilities available for control and influence are often limited to heuristics and approximate routines; in other words, there are no standing orders for Prussian drill but rather instructions for white water rafting. The new scientific

[8] "Everything should be made as simple as possible, but not simpler." It is not entirely clear, however, whether Einstein expressed himself precisely in that way. The closest reliable quotation is perhaps this: "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience." (From "On the Method of Theoretical Physics," the Herbert Spencer Lecture, Oxford, June 10, 1933.)

procedure emerging at this stage is massive modelling, simulation[9] based on high-power computing techniques. To use the method of simulation means to calculate the very many scenarios of a system's behaviour under specific circumstances and thus to come up with maps of the system's behaviour. Potentially, they will show areas where no reasonable forecast and control is possible, because of extremely sensitive dependence on some intractable influence factors. But one may also hope to find realms of nearly linear behaviour. These regions allow for secure influence on the system.

The procedure thus generates outcomes of a novel type. Instead of impressive numbers suggesting seriousness and certainty we end up with a more qualitative description of the phenomenon, describing its behaviour under various circumstances, aiming at identifying basins of robustness, or else pockets of order for the system. Forecasting climate changes will never result in figures with four decimal points, only in responsible estimations together with revealed margins of error. Formidable precision – once the very hallmark of science – may now indicate insincerity of the analysis. It will take some time to explain that to our customers!

All that will result in a new situation in science. How to keep the promise to pin nature down to a transparent mathematical description if such a strategy ruins the objective of investigation by oversimplifying the framework of analysis? How then can scientific results be made accessible to non-specialists? Science may find itself in a position where it has no choice but to breach the Renaissance pact.

Here is Sandra Mitchell's motivation for the swerve towards complexism[10]:

> The long-standing scientific and philosophical deference to reductive explanations founded on simple universal laws, linear causal models, and predict-and-act strategies fails to accommodate the kinds of knowledge that many contemporary sciences are providing about *the* world. (Mitchell 2009a, b, p. 118)

Does that mean objecting to the importance of fundamental laws in the sciences? Yes and no. Of course, there are fundamental laws in physical theories, and no doubt our present physical theories are the best theories of nature we have ever possessed. In that sense, fundamental laws are important. On the other hand, not much follows from the possibility of describing a human being as a relatively stable cloud of elementary particles. In that sense, I agree with Philip Anderson:

> The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. In fact, the more the elementary particle physicists tell us about the nature of the fundamental laws, the less relevance they seem to have to the very real problems of the rest of science, much less to those of society. (Anderson 1972, p. 394)

[9] "Simulation" refers to the process of constructing, using, and justifying a model that involves analytically intractable mathematics (Winsberg 2003, p. 105). At this stage, there is no established terminology, not to mention even clearly defined content for the new procedure.

[10] As defined above.

It seems then that we should accept Mitchell's conjecture that, generally, in the sciences, there are neither fundamental explanations nor any kind of universal investigative strategies. And this, I take it, is a positive message for science.[11]

References

Anderson, P.W. 1972. More is different. *Science* 177: 393–396.
Coates, J. 1996. *The claims of common sense. Moore, Wittgenstein, Keynes, and the social sciences*. Cambridge: Cambridge University Press.
Coates, J. 2012. *The hour between dog and wolf. Risk-taking, gut feelings and the biology of boom and bust*. London: Fourth Estate.
De Vreese, L. 2006. Causal pluralism and scientific knowledge: An underexposed problem. *Philosophica* 77: 125–150.
Dowe, P. 2000. *Physical causation*. Cambridge: Cambridge University Press.
Esfeld, M. 2012. *Philosophie der Physik*. Berlin: Suhrkamp Verlag.
Hagmann, P. 2005. *From diffusion MRI to brain connectomics.* Thesis, EPFL Lausanne.
Hitchcock, C. 2003. Of humean bondage. *The British Journal for the Philosophy of Science* 54(1): 1–25.
Keynes, J.M. 1921. *A treatise on probability. Collected writings*, vol. 8. Cambridge: Cambridge University Press, 1990.
Keynes, J.M. 1936. *General theory of employment, interest, and money. Collected writings*, vol. 7. Cambridge: Cambridge University Press, 1978.
Ladyman, J., J. Lambert, and K. Wiesner. 2011. What is a complex system? *Pittsburgh PhilSci Archive*.
Lipton, P. 2005. The truth about science. *Philosophical Transactions of the Royal Society B* 30: 1259–1269.
Mitchell, S.D. 2008. Taming causal complexity. In *Philosophical issues in psychiatry: Explanation, phenomenology, and nosology*, ed. K.S. Kendler and J. Parnas, 125–131. Baltimore: The Johns Hopkins University Press.
Mitchell, S.D. 2009a. *Unsimple truths. Science, complexity, and policy*. Chicago: Chicago University Press.
Mitchell, M. 2009b. *Complexity: A guided tour*. Oxford: Oxford University Press.
Murphy, D. 2008. Levels of explanation in psychiatry. In *Philosophical issues in psychiatry: Explanation, phenomenology, and nosology*, ed. K.S. Kendler and J. Parnas, 99–124. Baltimore: The Johns Hopkins University Press.
Persson, J. 2007. Mechanism-as-activity and the threat of polygenic effects. *Pittsburgh PhilSci Archive*.
Price, H. 2001. Causation in the special sciences: The case for pragmatism. In *Stochastic causality*, ed. M.C. Galavotti et al., 103–120. Stanford: CSLI Publications.
Price, H. 2007. Causal perspectivalism. In *Causation, physics, and the constitution of reality: Russell's republic revisited*, ed. H. Price and R. Corry, 250–292. Oxford: Clarendon Press.
Psillos, S. 2010. Causal pluralism. In *Worldviews, science and us. Studies of analytical metaphysics*, ed. R. Vanderbeeken and B. D'Hooghe, 131–151. Hackensack: World Scientific.

[11]Many thanks to Alan Fahy for improving the English and for the jazz metaphor, to Niels Dechow and to the members of my PhD course for debating the economic topics, and to Sebastian Urchs, Ireneusz Kojder and Felix Hasler for consultation about brains.

Seung, S. 2012. *Connectome. How the brain's wiring makes us who we are*. Boston: Houghton Mifflin Harcourt.

Urchs, M. 2001. Causal braids. On weakly transitive causality. In *Current issues in causation*, ed. M. Ledwig and W. Spohn, 151–162. Paderborn: Mentis.

Winsberg, E. 2003. Simulated experiments: Methodology for a virtual world. *Philosophy of Science* 70: 105–125.

Confessions of a Complexity Skeptic

Raphael Scholl

1 Introduction

Max Urchs argues in his contribution to the present volume that scientists and philosophers of science should be more mindful of complexity. In this he agrees with a number of recent contributions by such authors as Melanie Mitchell (2009a) and Sandra Mitchell (2009b). In her book *Unsimple Truths*, Sandra Mitchell argues that progress in scientific understanding will increasingly require complexity thinking, and that the philosophy of science will need to adjust its meta-reflections accordingly. Mitchell thinks that a host of issues will have to be reconsidered: reduction and emergence, lawfulness, scientific method, prediction, and policy analysis.

In his discussion of cases from economics and the neurosciences, Urchs touches on many of the same issues. In the comments in hand, it is not my goal to engage with the debate about complexity as a whole. Instead I will focus on objections in three of Urchs's main areas of discussion: economic markets, the changing legacy of the scientific revolution, and neuroscience. I am particularly interested in the question of what we can learn from individual cases from the sciences – that is, on the proper use of case studies in philosophy of science.

First, Urchs argues that complexity thinking is helpful in modern economics, and he identifies John Maynard Keynes as an early exponent of this view. I will argue that a close consideration of Keynes's science should not leave us unambiguously favorable to complexity thinking. Second, Urchs makes the historical claim that a certain kind of relationship between science and society was established during the scientific revolution, and that this will have to change to accommodate complexity thinking. I will argue that the historical claim is difficult to maintain, and that the

R. Scholl (✉)
History and Philosophy of Science, Institute of Philosophy, University of Bern, Sidlerstr. 5, CH-3012 Bern, Switzerland
e-mail: raphael.scholl@gmail.com

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_15,

changing relationship between science and society is, at the very least, still up for debate. Third, Urchs discusses the neurosciences as an area where complexity thinking will potentially be helpful. I neither affirm nor reject the thesis, but I argue that the neurosciences as a case study are not suitable for Urchs's purposes on strictly methodological grounds.

It will be useful for the rest of the discussion to have some idea of what is meant by complexity (for a recent discussion, see references cited above and Ladyman et al. 2012). There is no consensus answer, and Urchs's own approach to delineating the phenomenon is rather impressionistic. For the present discussion, I will think of complex systems as involving (a) a large number of entities, (b) a large number of possible (although perhaps simple) interactions, and (c) multiple relevant levels of description. The system at the aggregate level will display some sort of organization or adaptation, and it may moreover be subject to continuing evolution, such that the entities or their interactions change over time. I take it that these features roughly capture the systems that authors like Sandra Mitchell and Melanie Mitchell are talking about, and such systems would allow the occurrence of many of the phenomena that Urchs attributes to complex systems (among these are deterministic chaos, non-linearity, surprise[1] and unpredictability).

2 Keynes, Complexity, and the Ineffectiveness of Monetary Policy During the Economic Crisis

2.1 Keynes on Parts and Wholes

Urchs argues that John Maynard Keynes must be understood as a precursor of complexity thinking. He quotes from the preface of the French edition of *The General Theory of Employment, Interest and Money* of 1939:

> I argue that important mistakes have been made through extending to the system as a whole conclusions which have been correctly arrived at in respect of a part of it taken in isolation. (Keynes 1973a, p. xxxii)

Urchs wants us to read this in the context of complexity – perhaps as an endorsement of system-level emergence. But we must take a close look at what Keynes had in mind here. I will argue that Keynes can be reinterpreted in terms that do not unambiguously support complexity thinking.[2]

[1] Unlike Urchs, however, I would strongly caution against a definition of complexity in which subjective psychological experiences such as surprise play a role.

[2] Urchs also gives us quotations from Keynes's *Treatise on Probability* (1973b), and I have no comments to make about these. I agree that it would be worthwhile to read the *Treatise on Probability* from the point of view of complexity.

Keynes's goal in the French preface is to delineate the main distinguishing features of his approach to economic theory, and in particular the reasons why he speaks of a "general theory". Immediately before the quotation above, he writes:

> I have called my theory a *general* theory. I mean by this that I am chiefly concerned with the behaviour of the economic system as a whole, – with aggregate incomes, aggregate profits, aggregate output, aggregate employment, aggregate investment, aggregate saving rather than with the incomes, profits, output, employment, investment and saving of particular industries, firms or individuals. (p. xxxii)

This aggregate-level description laid the foundation for the current division of economics into micro- and macroeconomics. Keynes goes on to illustrate the approach using an example. His theory says that for the economy as a whole, savings must equal investment. Now, if taken as a statement about individual economic actors, this is plainly false, since there is no reason why an individual actor's investment should bear any relationship to his or her savings:

> Quite legitimately we regard an individual's income as independent of what he himself consumes and invests. But this, I have to point out, should not have led us to overlook the fact that the demand arising out of the consumption and investment of one individual is the source of the incomes of other individuals, so that incomes in general are not independent, quite the contrary, of the disposition of individuals to spend and invest; and since in turn the readiness of individuals to spend and invest depends on their incomes, a relationship is set up between aggregate savings and aggregate investment which can be very easily shown, beyond any possibility of reasonable dispute, to be of exact and necessary equality. (pp. xxxii–xxxiii)

Keynes regards this as a "banale conclusion" with interesting consequences. It follows, for example, that when faced with a large burden of debt, the rational course of action for an individual company is not the same as for an entire national economy. An individual company deals with debt by increasing the ratio of income to expenditures. However, if an entire national economy cuts its spending, then – because one person's spending is another person's income – everybody will be poorer off. Keynes writes:

> [I]t becomes evident that an increased propensity to save will *ceteris paribus* contract incomes and output; while an increased inducement to invest will expand them. (p. xxxiii)

That increased savings can lead to a loss of wealth is sometimes referred to as "the paradox of thrift". It constitutes the minimal core of recent disputes about how to deal with the economic crisis that started in 2008. Proponents of "stimulus" argue that cutting government spending would be counterproductive, since increased government savings will, as Keynes said, "contract incomes and output" of private actors like firms and households. Proponents of "austerity" argue that nations with too much debt must decrease spending to bring their budgets into balance, and that increased government spending would be ineffective stimulus.[3]

[3]Proponents of austerity would not, however, deny the savings/investment equality. Instead, they might argue that government investment would increase interest rates and so "crowd out" private sector investment. Or they might argue that increased government spending would be taken as a

Thus aware of the context of Keynes's statement about the properties of parts and wholes, should we read his remark in the context of the debates about complexity? I don't think so, for a much more pedestrian story suggests itself. It seems that Keynes's subdiscipline-founding breakthrough was to identify a successful aggregate-level (as opposed to component-level) description of causal factors in a national economy. His description in terms of aggregate incomes, outputs, profits, and so on, was successful in the sense that robust causal relationships could be identified among these aggregate causal factors, and in the sense that these causal relationships could be used to explain interesting phenomena (such as the behavior of economies in times of recession). The next section's more thorough discussion of the *IS-LM* model, which is generally taken as a formal model of Keynes's *General Theory*, will make this even clearer.

It is at least not immediately obvious how complexity enters the picture. The number of components in Keynes's picture is in fact a reduction relative to what one would expect: It turns out there can be meaningful economic models that completely disregard the fact that economies are in reality made up of hundreds of thousands of individual actors. Apparently, if one identifies the appropriate causal factors at the aggregate level, one can abstract away much of the "complexity" of an economy. Similarly, the number of interactions between components does not multiply in Keynes's description. There are two levels of description in play, but I assume that this by itself does not qualify a model for the complexity label. Perhaps there is some other sense of complexity which applies to Keynes's science – but Urchs's paper is scant help, since his definition of complexity is elusive.

2.2 The Ineffectiveness of Monetary Policy During the Economic Crisis

Another example discussed by Urchs suggests, upon close consideration, the same story: component-level complexity giving way to simpler aggregate-level models. Urchs reminds us that during the economic crisis, a standard remedy for slowing economic growth has proved ineffective: lowering interest rates in order to stimulate economic output. Urchs quotes John Coates:

> Economists assume economic agents act rationally, and thus respond to price signals such as interest rates, the price of money. In the event of a market crash, so the thinking goes, central banks need only lower interest rates to stimulate the buying of risky assets, which now offer relatively more attractive returns compared to the low interest rates on Treasury bonds. But central banks have met with very limited success in arresting the downward momentum of a collapsing market. One possible reason for this failure could be that the chronically high levels of cortisol among the banking community have powerful cognitive effects. (p. 209)

---

sign of higher government deficits in the future. This in turn would lead to an expectation of higher future corporate taxes, which would again induce companies in the present to reduce investment.

Perhaps the supposed relationship between the monetary base, interest rates and economic output is, as Urchs writes, "like a compass that will work properly only if there is no wind" (p. 208)?

The invocation of cortisol levels is certainly in the spirit of Sandra Mitchell's arguments that most interesting phenomena will ultimately have to be understood at a number of interacting levels. Her main example is clinical depression, where we must expect a full understanding to include both molecular factors *and* the patient's social environment, among other things. Similarly, economists may have to learn to model the behavior of bankers not just in terms of rational responses to interest rates, but also in terms of cortisol-related psychological processes of individuals.

Urchs himself expresses some skepticism about hormone-based explanations in economics. However, he is correct in noting that *if* macroeconomists must begin to model the hormonal state of individual traders, *then* macroeconomic models will increase in complexity in a number of possible senses: the number of interacting parts will increase, as will the potential for interactions among the parts and the required number of levels of description.

However, it is far from certain that irrationality-inducing cortisol levels are needed to explain the ineffectiveness of monetary policy during the economic crisis.[4] Some macroeconomists would argue that the ineffectiveness of monetary policy during the crisis is explained by aggregate-level factors – the kind that customarily enter into macroeconomic analyses. In particular, they would point to the *IS-LM* model, which is taken as a formal representation of Keynes's *General Theory*. I ask the reader to bear with me as I briefly introduce the *IS-LM* model. On this basis, it will then be possible to argue that the *IS-LM* model offers an analysis of the ineffectiveness of monetary policy during the economic crisis which is plausibly non-complex.

For a full presentation of the *IS-LM* model, I refer the reader elsewhere.[5] The basic and familiar idea is that economic output in an economy is determined both by the market for goods and services and by the market for money. More specifically, the model tells us how the interplay between the two markets determines interest rates and economic output. In a typical presentation of the *IS-LM* model (see Fig. 1), interest rates are indicated on the vertical axis and GDP is indicated on the horizontal axis. The *IS* curve (for "investment–saving") represents the market for goods and services and the *LM* curve (for "liquidity preference–money supply") represents the market for money.

Let us look at the *IS* curve first, where the interest rate is the independent variable and GDP is the dependent variable. Changes in the market for goods and services can increase or decrease economic output. Decreases in interest rates will make the

[4]This is not a rejection of neuroeconomics (for a discussion, see Mäki 2010). My rather more modest argument is that the neurosciences may not be needed in the particular case discussed by Urchs.

[5]See for instance Chaps. 10 and 11 in the textbook by N. Gregory Mankiw (2002), and for a brief blog-sized overview see Paul Krugman (2011).

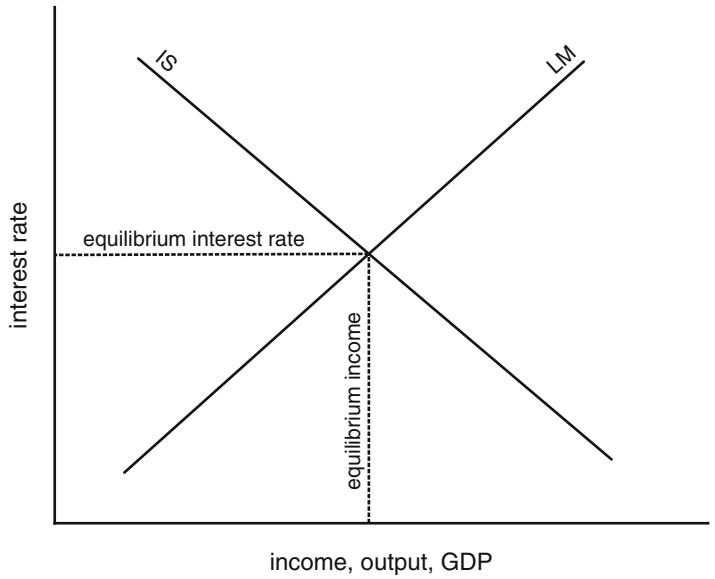

Fig. 1 The *IS-LM* model after Mankiw (2002) and Krugman (2011). See text for explanations

financing of investments less costly, and so lower interest rates increase economic output (this is why the *IS* curve is downward sloping). Several factors can shift the *IS* curve. Increased government purchases will increase overall expenditures, overall income, and economic output. Similarly, a tax decrease will also expand expenditure and income. Both sorts of changes would shift the *IS* curve to the right. Conversely, a decrease in government spending or a tax increase would shift the *IS* curve to the left.

Let us now turn to the *LM* curve, where GDP is the independent and interest rate the dependent variable. A higher GDP will mean higher demand for money, since more interactions requiring money take place, and so interest rates will be higher if GDP is higher (this is why the *LM* curve is upward sloping). Again there are factors that can shift the *LM* curve. Interest rate at any given GDP can be affected by changes in the amount of money available, that is, by the monetary policy of central banks. An increase in the monetary base will tend to reduce interest rates: it will shift the *LM* curve to the right. The converse happens if the monetary base is reduced.

It is important to understand that the *IS* and *LM* curves represent possible *equilibrium points* in the market for goods and services and the market for money. Let's look at the market for goods and services first. If GDP were above the level compatible with a certain interest rate, this would mean that companies are producing more goods than are demanded by current investment. Hence, companies would accumulate inventory and in response decrease production, which would bring GDP in line with interest rates. Conversely, if GDP were below the level compatible with the interest rate, more goods would be demanded than produced,

inventories would decline, and companies would increase production. This would again bring actual GDP in line with the interest rate. In the market for money, an equilibrium process is also at work. If interest rates were above equilibrium levels, individuals would try to convert more of their money into interest-bearing bank deposits or bonds, and this excess supply of money would cause banks and bond issuers to lower the interest rates they offer. Conversely, if the interest rate were below equilibrium, banks and bond issuers would offer higher interest rates to attract scarce money. The economy as a whole is at the point where both the market for goods and services and the market for money are in equilibrium: that is, at the intersection of the *IS* and *LM* curves.

Now how can the *IS-LM* model explain the ineffectiveness of monetary policy during the crisis? This is where I take it that there is controversy within the science of economics. One explanation on offer, defended forcefully by the Nobel Prize winner Paul Krugman, is that the economy in the crisis was in a so-called "liquidity trap".[6] The basic idea is that many individuals in the economic crisis suddenly found themselves in a position where they had a debt burden to reduce. This caused expenditures to go down drastically – it shifted the *IS* curve far to the left, reducing interest rates and economic output. Central banks reacted to the slump in output by increasing the money supply (shifting the *LM* curve to the right). In normal times, this would reduce interest rates and stimulate investment spending. However, since the shift in the *IS* curve had already brought interest rates close to zero, a shift in the *LM* curve could not affect them further, and so monetary policy had to remain ineffective. This situation is shown in Fig. 2.

The liquidity trap offers an explanation for the failure of monetary policy during the great recession which impresses by its relative simplicity. It is not within my competence to say whether this explanation actually is correct. But if it is correct, then macroeconomic success was again achieved by modeling the correct aggregate factors and the correct causal relationships between them. And as far as I can tell, the *IS-LM* model is not complex in the sense discussed here. Both in terms of entities and interactions, the model is relatively parsimonious.

We must ask, however, whether the explanation provided by the *IS-LM* model is truly independent of an explanation which invokes, for instance, the cortisol level of individuals. Aren't neuroscientific factors simply *what explains* that the *IS* curve shifted far to the left – or in other words, isn't the *IS-LM* analysis a mere description of a process which Urchs (quoting Coates) wants to explain causally by appeal to cortisol levels?

In a trivial way, neuroscientific facts are certainly part of the complete causal story. The shift of the *IS* curve is explained by a reduction in overall planned expenditure in the economy, caused by many economic actors suddenly facing a debt burden. All of this must be realized at the individual and neuronal level – and not all individuals will face the same sort of debt, or react to it psychologically and neurally in the same way. So micro-level realizations of the aggregate-level fact of

[6] See also Mankiw's discussion of the liquidity trap in his textbook (Mankiw 2002), Chap. 11.

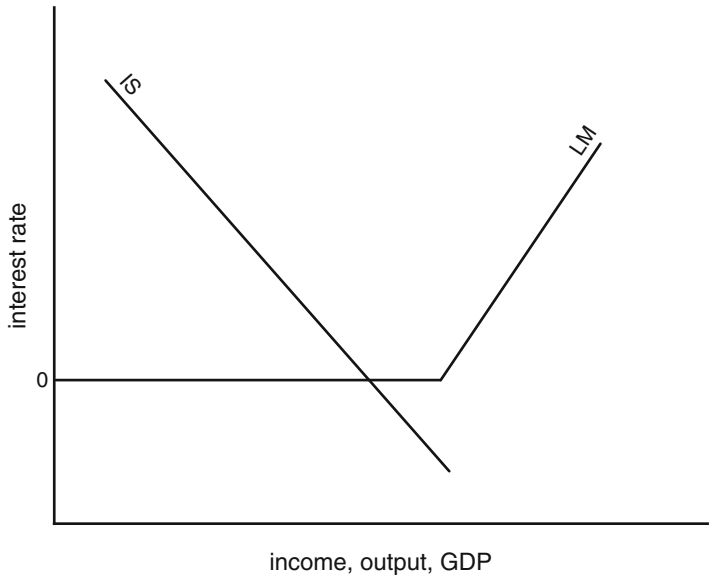

Fig. 2 The *IS-LM* model in a "liquidity trap" (after Krugman 2011). See text for explanations

"a large decrease in planned expenditure" certainly exist and are multiply realized. But I presume that this would be largely uncontested.

The more pressing question is whether the cortisol-levels explanation is intended as *complementary* or *alternative* to the customary macroeconomic analysis. Returning to the first sentence of Urchs's quote from Coates, we read:

> Economists assume economic agents *act rationally*, and thus respond to price signals such as interest rates, the price of money. (p. 209, my emphasis)

Since economic agents during the crisis seemed not to react to price signals, cortisol levels are offered as an explanation of this irrationality. They are thus clearly intended as an *alternative* explanation: The neuroscientific facts are taken to explain something that on a customary analysis remains obscure. However, on the *IS-LM* analysis economic agents were not behaving irrationally, and so there is nothing for the cortisol-levels hypothesis to explain! With interest rates having approached the zero lower bound, an increase in the monetary base could not have been expected to alter interest rates further – even if all economic agents were behaving rationally. Cortisol levels and other neuroscientific facts surely constitute part of the micro-level realization of macroeconomic states, but they are mere intermediate causes: the explanatory work is done by the aggregate-level theory.

Friends of complexity might accept the foregoing analysis but insist that the explanation of the ineffectiveness of monetary policy during the crisis is *precisely* what complexity is all about. Among Urchs's features of complexity is:

> The very mechanism which generates a system's behaviour is changing, leading thereby to the evolution of new forms of connections. (p. 204)

One might tell the story thus: in normal times, the economic system's behavior is such that an increase in the monetary base increases economic output. In the crisis, the mechanism generating this behavior changed, leading to unexpected new results. But I think it would be unwise to use the term "complexity" in case where we are simply dealing with misconceptions about the mechanisms operating in the system. The relevant macroeconomic model was perhaps not widely understood, but it is not particularly complex.

Urchs writes that standard models in the economic crisis were somewhat like a compass that only works if there is no storm. It is true that the expected relationship between expansion of the monetary base and economic output broke down during the crisis, but the actual compass – the *IS-LM* model – may have weathered the storm successfully. The *IS-LM* model is certainly *more* complex than the basic expectations about monetary policy and economic output, but it is still relatively simple.

It is of course possible to criticize the *IS-LM* analysis from the point of view of complexity. For example, W. Brian Arthur (1999) argues that equilibrium analyses are generally in need of expansion. I am sympathetic to this. But the case for such expansions must rest on phenomena that the equilibrium analysis *cannot* explain. And if my account above is correct, then the ineffectiveness of monetary policy during the crisis of 2008 is not such a phenomenon.

To summarize, I remain skeptical that either Keynes's talk about the properties of parts and wholes or the ineffectiveness of monetary policy during the crisis offer particularly good examples of complexity. Macroeconomic success in both cases is not obviously dependent on an understanding of what Sandra Mitchell or Max Urchs would call complexity. To the contrary, it seems that the key was to find an aggregate-level description of the economy which could successfully handle the system's component-level diversity. All of this leaves open the possibility, of course, that other cases from economics do in fact necessitate complexity thinking.

3 The "Renaissance Pact"

Urchs claims that science and society entered, some 400 years ago, into a "Renaissance pact" or "promise":

> We, the scientists, will work hard to uncover the most fundamental structures of the world. We will describe them in mathematical language in a Golden Book of Nature. Everyman receives a precise and perspicuous picture of reality that lays the foundation for technological progress. We will be rewarded for that. (p. 215)

He claims that the pact is now in need of revision because of complexity.

It is not clear how this Renaissance pact should be interpreted. Urchs paints the scientific revolution as a response to an earlier breakdown in the communication between science and society, largely because science was theoretical rather than applied. From this he develops the notion of a new relationship between science and

society from the renaissance onward – the aforementioned pact or promise. Urchs does not refer to the historical literature to back up these claims. To the best of my knowledge, they are untenable.

One difficulty of the proposal is that it may not even be a testable historical thesis. This is because it is hard to conceptualize. Pacts have signatories, and promises are made by someone to someone. Yet who would that be? Even if we take all early modern scientists to pursue roughly the goal stated above (to "uncover the most fundamental structures of the world"), which individuals or groups from society offered a reward in return? Was it because a reward was promised that science was professionalized and eventually publicly funded, some two centuries later? As a matter of historical scholarship, I do not know how this thesis can be made precise or plausible.

But I suspect that the idea must be taken metaphorically: Over the past four centuries, science has increasingly served society in a certain way, and now this may be changing. Leaving the problematic historical claims aside, Urchs may be saying that the epistemological goals of science may need to be adjusted.

Urchs seems to be thinking along the lines of Sandra Mitchell's Chap. 5, which discusses policy in a world of complexity. On one conception, the task of science is to make definitive predictions about the consequences of particular courses of action (for instance, on whether anthropogenic greenhouse emissions will or won't cause radical climate change), and these predictions then allow us to act so as to achieve our desired outcomes. However, from a science of complexity we may not get definitive predictions, but only an analysis of different possible outcomes. Parts of the outcome space may show such sensitive dependence on initial conditions that no predictions are possible. Policy in turn may not be able to choose one outcome as a goal: we may have to settle for policy which is *robust* in the sense that it leads to acceptable outcomes in a variety of different scenarios, and our policy may in any case have to adapt itself as more is learned about the systems with which we are dealing.

I find these ideas by and large unobjectionable, and to the extend that Urchs is thinking about the same sorts of things, I agree with him. But we would have to think carefully about whether any of this leads to a change in the relationship between science and society. I rather suspect, for instance, that Mitchell gives us a somewhat caricatured view of the "predict and act" model on which science is supposed to have acted in the past. And the strength of the thesis clearly depends on describing current practice correctly: if the new principles urged by the friends of complexity are already part of science, then the new approach to prediction and policy may not differ much from business as usual.[7]

[7] Another worrisome note is offered by a reviewer of Mitchell's book. Kristin Shrader-Frechette (2013) points out that some of Mitchell's phrases such as "flexible management", "continued investigation" and "learn by doing" can be weaponized by Washington lobbyists to mean toothless regulation, delayed action and science-blindness.

In summary, it is not entirely clear what Urchs's aims are in formulating the thesis of the "Renaissance promise". As a historical thesis it fails, and as a metaphor its precise content remains to be articulated.

4 On the Function of Case Studies in Philosophy of Science

Philosophy of science works best with a strong empirical component, and so the use of case studies from historical or contemporary science is necessary and welcome. But it is a challenge to choose the right cases for any given philosophical purpose. To illustrate the point, suppose we are interested in heuristics for the generation of novel scientific hypotheses, and part of our argument is to show that our philosophical proposal operates in a case study from actual science. It would surely tell us little to see that one or another method of theory generation applies to a *trivial* scientific discovery: it will not come as a surprise to anybody that some simple heuristics can illuminate successful but modest extrapolations from known and well confirmed theories. In order for our cases to *test* the value of the proposed heuristics, they must be scientific theories of acknowledged novelty and originality. In this case, hypotheses that led to a Nobel Prize are perhaps promising candidates. In general, we may say that we must challenge our philosophy with *hard cases*.

The problem of the choice of case studies is particularly acute in the complexity debate. One use for case studies is to show that we can solve certain problems only by being mindful of one feature of complexity or another – for example, by studying the interactions of a great many components, or by integrating causality at multiple levels of organization or in multiple domains of investigation. But of course this argument can only be made once at least some such problems are actually solved. Only when some results are in can we analyze them and show how complexity thinking was epistemologically fruitful.

Urchs's section on neuroscience, however, is a mere promissory note: neuroscience *may* solve some or even many of its outstanding questions through methods that are more mindful of complexity. This is plausible, since the number of entities is known to be large and multiple levels of organization seem to be relevant. But little is accomplished by pointing to a subject which is at present insufficiently understood. Our insufficient understanding may stem from the fact that we have not been mindful of complexity – but we may also lack the right kind of mathematical methods, or perhaps the correct aggregate-level causal factors have not yet been identified (as was essential for Keynes's work). Without additional arguments for the use of complexity thinking, we do not learn much from neuroscience's outstanding questions beyond the fact that these questions exist.

A partial model for how to navigate this territory is Sandra Mitchell's *Unsimple Truths* (2009b). Discussing older concepts of emergence in her second chapter, Mitchell notes that philosophers in the nineteenth century spoke of emergence in the context of chemical properties (the fluidity of water, for instance) and biological properties (inheritance) which later received reductive explanations. Mitchell then

argues for what she hopes will be a more useful new concept of emergence. By opening her discussion with such philosophical precursors, Mitchell to some extent diffuses the fears of readers like me: she knows that the number of phenomena that seem intractable by customary methods has decreased for centuries, and that simply pointing to an as yet ill-understood area of science is no argument for the necessity of new methods. What is required in addition is an argument for why the remaining unsolved questions are different. Similarly, Mitchell discusses the case of major depressive disorder in some detail to show how complexity thinking has been required for understanding it. One can argue about whether Mitchell is entirely successful in her project. In particular, I wished for a much more detailed treatment of the case study. But at least her plan of attack is exactly right.

In summary, while the use of case studies in philosophy of science is necessary and welcome, it is also difficult. Cases must be chosen such that a true test of a philosophical thesis is possible. I argued that this does not succeed in Urchs's section on neuroscience (the case) and complexity (the philosophical thesis).

5 Conclusions

Having titled my comments "confessions of a complexity skeptic", I conclude by stressing – call it a hedge, if you must – that I have considerable sympathy for the project of the friends of complexity. I wish the debate a long and healthy life. It is difficult to approach philosophy of science from the perspective of the biomedical sciences without suspecting that the notion of complexity may be both analytically tractable and useful. Complex systems may lead us to a more robust concept of emergence, or to a better understanding of causal inference in biological and social systems. So my position should probably be described as *local skepticism*: I have made explicit concrete problems faced by Urchs's discussion of complexity, and this leaves the door wide open for others to make the case for complexity more compelling. However, the objections raised in my discussion have some claim to generality. In order to argue for or against complexity, we will need to consider cases from actual science very closely, and to think hard about what they can and cannot tell us.

References

Arthur, W.B. 1999. Complexity and the economy. *Science* 284: 107–109.

Keynes, J.M. 1973a. *The general theory of employment, interest and money*, The collected writings of John Maynard Keynes, vol. VII. London: Macmillan.

Keynes, J.M. 1973b. *Treatise on probability*, The collected writings of John Maynard Keynes, vol. VIII. London: Macmillan.

Krugman, P. 2011. IS-LMentary. http://krugman.blogs.nytimes.com/2011/10/09/is-lmentary/. Accessed 16 Apr 2013.

Ladyman, J., J. Lambert, and K. Wiesner 2012. What is a complex system? *European Journal for Philosophy of Science* 3(1): 33–67.

Mäki, U. 2010. When economics meets neuroscience: Hype and hope. *Journal of Economic Methodology* 17(2): 107–117.

Mankiw, N.G. 2002. *Macroeconomics*. 5th ed. New York: Worth Publishers.

Mitchell, M. 2009a. *Complexity: A guided tour*. New York: Oxford University Press.

Mitchell, S.D. 2009b. *Unsimple truths: Science, complexity, and policy*. Chicago: University of Chicago Press.

Shrader-Frechette, K. 2013. Sandra D. Mitchell: Unsimple truths: Science, complexity, and policy. *The British Journal for the Philosophy of Science* 64: 449–453.

New Directions in the Philosophy of Biology: A New Taxonomy of Functions

Cristian Saborido

1 Introduction

Many things in the philosophy of science have changed in the last decades. A clear example of this is the debate on the concept of biological function. The supposed teleological and normative character of so-called functional explanations is at the heart of one of the most profitable and valuable discussions that currently exist.[1] What is more, the development of this discussion allows us to understand many of the changes and controversies that have marked the direction that the philosophy of biology has taken at the beginning of the twenty-first century.

In this paper, I review the different theories on functional explanation that can be found in the current debate in philosophy of biology. I take it that the current state of the philosophical discussion is dominated by two major classical perspectives that address functional explanations and consider functions as a kind of disposition. The first of these major views, the "causal-role" or "systemic" approach, describes functions as causal effects of a biological trait in the frame of a system or organism. The second approach, the "evolutionary" one, considers that functions can be identified with the biological effects that are the "causes of existence" of biological traits by appealing to the evolutionary role of these effects. Although these two views have a relatively long history, in the last years a number of new theories have emerged within these perspectives. In the taxonomy of functions I present here, I introduce these new theoretical formulations of the causal role, as well as explain the evolutionary approaches and compare and critically analyze their strategies and explanatory focus.

[1] See, for instance, the following collections: Allen et al. 1998; Buller 1999a; Ariew et al. 2002; Krohs and Kroes 2009.

C. Saborido (✉)
Department of Logic, History and Philosophy of Science, National Distance Education University (UNED), Madrid, Spain
e-mail: cristian.saborido@fsof.uned.es

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_16,

I will begin by presenting a critical review of these two ways of interpreting the notion of function in light of the current theoretical proposals. I will then analyze an attempt to overcome this dichotomy: the recent Organizational Approach. I claim that this last approach constitutes the major novelty in the philosophical discussion on functions. According to organizational theories, a function is a disposition of a particular current biological trait that has explanatory relevance, in organizational terms, with regard to the presence of the function-bearing trait. The organizational account claims that a functional effect can be understood as a condition of existence of that very trait (without appealing to evolutionary history) to the extent that it is a necessary condition for the process of biological self-maintenance of the organism (see also Schlosser 1998 and McLaughlin 2001). In the present article, I maintain that the Organizational Approach implies an integration of the etiological explanatory strategy and the causal-role framework by considering that a function is both a cause for the existence and a current disposition of a biological trait token.

2 Functions, Teleology and Normativity

"Function" is a key notion in the biomedical sciences. In a general sense, every function is a disposition of a biological trait. In fact, it is quite usual to interpret functions in terms of dispositions, or related notions such as "powers", "abilities" or "potencies". Therefore, the functions of a trait would be determined by the potential causal effects of such trait under some given circumstances. According to many authors (see Popper 1959; Shoemaker 1980; Bird 2007) dispositions can be understood as nomic or causal roles, and this is precisely the way in which the different theories interpret the notion of function.[2] A functional trait is a trait that has a disposition to produce a specific effect that has relevance with respect to a goal (the achievement of a systemic capacity, the increment of the system's fitness, the preservation of the biological self-maintenance . . .). Therefore, all theories understand that functions are a kind of *dispositional* and *causal* effect.

However, there seems to be a general consensus that not every disposition is a function. Many authors defend that what characterizes functions is that these have a *normative* and *teleological* dimension.

Functions are teleological at least in one sense but they can also be so in another one. First of all, functional attributions imply a teleology because they seem to refer

[2]According to the classical definition based on the Simple Conditional Analysis, an item is disposed to do something in given circumstances if and only if this item would do that very same thing in the cases that these circumstances are present. So, for instance and following the canonical example proposed by Carnap: "*x* is soluble iff, when *x* is put into water, it dissolves". In this paper I claim that functions are dispositions in this sense. A trait T has a biological function F if and only if T has the disposition to perform F, or in other words, F is a function of T iff, given the "appropriate circumstances", T effectively performs F. Of course, a theory of functions should clarify what are the "appropriate circumstances" because, in the absence of a developed theory, a biological trait has a potentially undetermined list of potential effects or dispositions that can be interpreted as functions.

to certain "raisons d'être", "purposes" or "intentions" related to the entities to whom one attributes these functions. An effect of a feature is a function only in relation to an (internal or external) purpose to which this effect contributes.

Second, certain types of functional explanations are also teleological in a stronger sense, since they try to explain the existence of a feature through some effects or consequences of its own activity, i.e., its function(s). To affirm – by quoting what is probably the most recurring example in this debate – that "the heart's function is to pump blood" is, ultimately, equivalent to saying that this effect of the heart, the pumping of blood, is relevant in order to explain the existence, structure and morphology of hearts (Buller 1999b, pp. 1–7).

Therefore, it is possible to hold that there are two ways of talking about teleology. First, there is a teleology related to all those statements that refer to certain "ends", "goals" or "intentions", as in the case of functional explanations. And, second, there is a special kind of functional explanation according to which a system's trait having a certain function implies that there is an effect of that trait that explains the existence of the very trait in that system. These explanations are teleological because they offer an explanation (*logos*) about the existence of a specific feature precisely through the functional purpose (*teIos*) that we attribute to it. In Walsh's terms: "Teleology is a mode of explanation in which the presence, occurrence, or nature of some phenomenon is explained by appeal to the goal or end to which it contributes" (Walsh 2008, p. 103). It is this strong sense of teleology that defines many theories on functions, such as the etiological approach since its beginning (Wright 1973), and it is this interpretation of teleology that has been strongly criticized by many theorists from the so-called "non-teleological" perspectives (for example, Cummins 2002; Davies 2001).

In addition to possessing this teleological character, the concept of function is inherently normative to the extent to which it refers to some effect that is supposed to take place (Price 1995, 2001, pp. 12–15; Hardcastle 2002, p. 144). When a function is attributed, a certain rule is postulated at the same time, a rule which is applicable to the behavior of what we consider as functional. As McLaughlin (2001, 2009) has pointed out, functions show a particular type of relation between certain means and goals in a system, which go beyond the standard concept of causality and have a normative flavor: in order for some systemic goals to happen, some effects need to occur, effects to which we refer as functions. The attribution of functions consequently implies the postulation of a specific type of effect for the functional traits. This type-token relation is what allows us to evaluate a system's activity in normative terms. For example, saying that the heart's function is pumping blood is equivalent to affirming that tokens of the type "heart" *should* pump blood. In case of not doing so, the heart would not be working properly, i.e., according to a norm ascribed to tokens of the type "heart".

Clearly, the normative dimension of functions requires an appropriate theoretical justification of the criteria under which the functional relations are identified as such and distinguished from all the other causal relations in the activity of a system. Functions are understood as the norms that must be satisfied and it must be explained why this is so in order to defend that these causal relations must be accomplished whereas others (the non functional or "accidental" effects) simply occur.

Both aspects, teleology and normativity, overcome the traditional scheme of the causal classic explanations and therefore they mean a real challenge for a naturalistic perspective in science and philosophy (Achinstein 1977; Buller 1999b; Mossio et al. 2009). In the current debate many different theories have faced this challenge from very different perspectives. In the following, I will review the main current theoretical analyses of the concept of function in Biology, which are mainly classified under the perspectives of causal role and evolutionary approaches. After that, I will introduce the organizational view and explain its approach to these teleological and normative dimensions.

3 Functions as Causal Roles

One of the classic strategies consists in interpreting functional dispositions as causal roles of a specific trait with respect to a capacity or activity of the global system. A disposition is a function if and only if it has a causal effect that contributes to the achievement of a higher-level activity or goal. The discrepancies here are related to the different ways of grounding the systemic notion of "goal".

The most comprehensive theory of biological functions is the "systemic account" (SA), first presented by Cummins in his paper "Functions" (Cummins 1975). Cummins claims that functions are contributions of certain parts or processes to the achievement of some systemic goal. According to this approach, functions are causal effects or dispositions of a trait, i.e., means–end relations contributing to some distinctive capacity of the global system.

Therefore, and contrary to the interpretation of many other approaches, Cummins' interpretation holds that functions have no explanatory power regarding the existence of the functional trait. In Cummins words:

> Teleological explanations and functional analyses have different *explananda*. The *explanandum* of a teleological explanation is the existence or presence of the object of the functional attribution: the eye has a lens because the lens has the function of focusing the image on the retina. Functional analysis instead seeks to explain the capacities of the system containing the object of functional attribution. Attribution of the function of focusing light is supposed to help us understand how the eye, and, ultimately, the visual system, works. In the context of functional analysis, a what-is-it-for question is construed as a question about the contribution 'it' makes to the capacities of some containing system. (Cummins 2002, p. 158)

For Cummins, this teleological interpretation of functions is mistaken. It is a vestige of a pre-scientific conception of nature that cannot take place in a naturalistic theory.[3] By rejecting the teleological dimension, Cummins' conception has the consequence of lacking definite criteria to determine which systemic capacity is the legitimate goal for a functional ascription. At most, this theory can say that a function is a contribution to a systemic capacity that is determined by the pragmatic

[3]In a recent work, P.S. Davies defends that the act of considering functional ascriptions as teleological and normative corresponds to a "conceptual conservationism", with psychological and cultural roots, which should be avoided to build proper, objective knowledge (Davies 2009).

interests of the researcher (Cummins 1975, p. 759).[4] Cummins' approach considers that functions refer to current relations between parts and capacities in a wide range of systems and consequently dissolves the problem of teleology of functions by reducing these to any causal contribution to a systemic capacity.

This approach argues that any trait's effect can be considered as a function if it is a contribution to a systemic capacity and consequently it provides no clear theoretical grounds to distinguish between the notions of "function" and "effect". This is the reason why many authors have argued that the causal role account is "too liberal" (Cfr. Davies 2001, pp. 73–75). According to this criticism, Cummins does not provide any clear criteria for identifying the relevant systemic goal or capacity. Once the teleological dimension of functions had been left aside by the causal role account, other dispositional approaches appeared which focused on providing criteria which are naturalized, i.e., grounded in some constitutive features of the system and not related to an extrinsic evaluative decision of the observer and appropriate, i.e., in accordance with both scientific and everyday usage, to identify what counts as a target capacity of a functional relationship, from which the legitimate norms could be deduced. The different causal role approaches have proposed various criteria to identify these target capacities.

All this led to new theoretical formulations of this "systemic approach" (SA), directly derived from Cummins work, which tried to defend a more sophisticate definition of the notion of function. These formulations restrict functional ascriptions to behaviors of parts of *hierarchically organized systems* (Davies 2001; Craver 2001). This version of SA fits the following definition:

A is a valid functional ascription for a systemic item I of a system S iff:

(i) I is capable of doing F,
(ii) A appropriately and adequately accounts for S's capacity to C in terms of *the organized structural or interactive capacities of components at some lower level of organization*,
(iii) I is among the lower-level components cited in A that structurally or interactively contribute to the exercise of C,
(iv) A accounts for S's capacity to C, in part, by appealing to the capacity of I to F,
(v) A specifies the physical mechanisms in S that instantiate the systemic capacities itemized. (Davies 2001, p. 89. Emphasis added)

Consequently, by restricting functions to hierarchically organized systems, the new SA approaches attempt to offer criteria for differentiating between every

[4]Cummins' analysis can be understood as an epistemological proposal: a functional analysis is interesting when the analyzed system has a remarkable organizational complexity. Thus, Cummins specifies three necessary conditions for this functional analysis:

(a) The analyzing capacities are "less sophisticated" than the analyzed capacity;
(b) The analyzing capacities are "different in type" from the analyzed capacity;
(c) The analyzing capacities exhibit a "complex organisation" such that together they explain the emergence of the analyzed capacity (Cummins 1975, p. 759)

potential effect of a trait and the real function of this trait, thus avoiding the "liberality problem" of Cummins definition. However, there is still a problematic characteristic in this SA strategy: it under-specifies functional ascriptions (Wouters 2005). The different formulations of the SA, even when only considering hierarchical systems, are not restrictive enough to offer a specific definition of function. In fact, at least three problems arise with the SA interpretation of functionality. First, there is not a principled criterion to distinguish between systems whose parts have functions and systems whose parts do not (Bigelow and Pargetter 1987; Millikan 1989). There are many examples of non-biological hierarchically organized systems whose parts are not subjects of functional ascriptions. Second, the SA is not able to adequately distinguish between functional contributions and dysfunctional or irrelevant effects (Millikan 1989; Neander 1991). The normative dimension of functions is missing in this approach. And third, the SA does not draw an appropriate distinction between effects that contribute to the achievement of a systemic goal in a "proper" functional way and accidentally useful effects, and consequently, the important distinction between "function" and "accident" is not grounded (Millikan 1993, 2002).

These fundamental weaknesses of the SA are precisely what the "Goal Contribution Approach" (GCA) has attempted to solve. This approach links the concept of function to the idea of goal-directedness. Accordingly, this approach introduces in the systemic framework more specific restrictions on what makes causal relations functions. Thus, according to the GCA, a function is a causal contribution to any (higher-level) capacity that constitutes a "goal state" of the analyzed system (Adams 1979; Boorse 1976, 2002).

The theorists of the GCA adopt a cybernetic definition of "goal-directedness" (Rosenblueth et al. 1943; Sommerhoff 1950). In Boorse's terms:

> A system S is 'directively organized' or 'goal directed' toward a result G when, through some range of environmental variations, the system is disposed to vary its behavior in whatever way is required to maintain G as a result. Such a system, it is said, shows 'plasticity' and 'persistence' in reaching G: when one path to G is blocked, another is available and employed. (Boorse 2002, p. 69)

This cybernetic characterization of "systemic goal" allows them to identify the goal states of a system in a naturalized and non-arbitrary way. In particular, theorists of the GCA describe biological systems as systems whose behavior is internally directed to achieving survival and reproduction and, accordingly, biological functions would be the internally generated contributions to these goals.

This perspective substantiates the causal relationship involved in functional behaviors, but at the cost of introducing norms whose application is, in fact, not restricted to the relevant kinds of systems and capacities. Cybernetic criteria may interpret dysfunctional behaviors of goal-directed systems as functional (cfr. Bedau 1992; Melander 1997). Every internal regulation leads the system to a concrete state that can be interpreted as a goal in cybernetic terms, independently of any other considerations as, for instance, the relevance or implications of achieving this state for the systemic viability. For example, a mammal that, due to a defect in its regulatory system, tends to maintain a constant fever can be considered as a

system cybernetically directed to this state (fever) and its fever should be interpreted as a systemic goal. The same can be said of many other cases of poor or wrong regulation, such as the case of autoimmune diseases. Identifying regulation with normativity involves considering many goal-directed states as legitimate systemic goals relevant to the theoretical grounding of functional ascriptions. Accordingly, the GCA still seems to under-specify functional attributions, and in some cases it appears to be an even less satisfactory account than the SA.

To sum up, I conclude that the causal-role approach, both in its SA and in GCA formulations, defines functions as current means–end relationships, and more specifically as current contributions of components to the emergence of a specific capacity of the containing system. Therefore, according to this view, functions are not teleological to the extent that they do not refer to any causal process that would explain the existence of the function bearer. I claim that this interpretation has many virtues for the scientific practice but in the end fails to provide a fully satisfactory ground for the normativity of functional attributions because it underdetermines the conditions for functional ascriptions. Causal role definitions turn out to be systematically under-specified: they do not restrict functional ascriptions to the relevant classes of systems and/or capacities.

4 Functions as Evolutionary Causes of Existence

As I have explained, the causal role approaches reject the teleological dimension of the notion of function and are incapable of providing definitive criteria for the theoretical grounding of the normativity involved in the notion of function. Significantly, this teleological character is precisely what enables us to account for the normative dimension of functional explanations. By restricting the term "function" only to those dispositions that explain the presence of the trait, teleological theories offer a clear criterion to determine the goal to which functional traits must conform. According to these theories, a function would be, ultimately, a disposition to contribute to the existence of the functional trait, and this contribution to the perpetuation of the trait is also the norm of its functioning. This teleological notion is at the basis of the mainstream approach in the current debate on functions: the Evolutionary Approach.

Most of the existing literature has favored this view, according to which an adequate understanding of functional attributions has to deal with the problem of teleology. In particular, both the teleological and normative dimensions are conceived of as being inherently related to the evolutionary role of the functional trait. Within this Evolutionary Approach, the most predominant view is the etiological approach (Wright 1973; Millikan 1989; Neander 1991; Godfrey-Smith 1994). These evolutionary-etiologic theories identify functions with the *causes of existence* of the functional trait. The etiological approach defines a trait's function in terms of its etiology (i.e., its causal history): the functions of a trait are past effects of that trait that causally explain its current presence.

In order to ground the teleological dimension of functions without adopting an unacceptable interpretation of the causal loop described by Wright, mainstream evolutionary accounts, usually called "Selected Effect Theories" (SET), have appealed to natural selection as the causal process that would adequately explain the existence of the function bearer by referring to its effects. In fact, according to the SET, functional processes are not produced by the same tokens of which they are supposed to explain the existence. Instead, the function of a trait is to produce the effects for which past occurrences of that trait were selected by natural selection (Godfrey-Smith 1994; Millikan 1989; Neander 1991). Selection explains the existence of the current functional trait because the effect of the activity of previous occurrences of the trait gave the bearer a selective advantage.

The main consequence of this explanatory line is its historical stance: what makes a process functional is not the fact that it contributes in some way to a present capacity of the system, but that it has the right sort of selective history. By interpreting functions as selected effects, the SET is able not only to deal with the problem of teleology, but also to ground the normativity of functions. SET identifies the norms of functions with their evolutionary conditions of existence: the function of a trait is to produce a given effect because, otherwise, the trait would not have been selected, and therefore would not exist. To contribute to the existence of the functional trait, through natural selection, is the functional norm.

One of the weaknesses of the SET is that natural selection cannot guarantee functionality to structures that have been selected in a historic moment for a certain reason and have been selected again for another distinct reason at another later time. These situations are very frequent in biology and have led writers like Gould and Vrba (1982) to propose the term "exaptations" for them.[5]

Another problematic point of SET is the presupposition that a trait is always selected against other alternatives because of its achievement of a concrete effect (i.e., the function). As some authors claim, this consideration is too strong. A biological trait can be interpreted as functional from an evolutionary stance simply because this trait contributed to the fitness of the organism and, consequently, to the perpetuation of this kind of organism and of the trait itself. The trait cannot be considered independently of the whole organism.

[5]To avoid this problem, many evolutionary theories specify that this selection has to occur in a period which is relevant for the current activity of the trait. Theorists such as Godfrey-Smith (1994), Griffiths (1993) and Schwartz (1999) have introduce new temporal restrictions to the SET. The so-called Modern History Theories consider that only recent history is relevant for functional ascriptions. According to this approach, the function of a trait is the effect that has caused this trait to be selected in the most recent period of time. Thus, it does not matter that bones had a metabolic function during a certain evolutionary period, because the reason for the current presence of bones is that they support the body, and that is now their proper function.

This kind of theory is able to account for the cases of exaptations, and includes the contribution to the fitness of the system as a condition to consider a concrete effect as a function. However, even when restricting the period of time, many of the objections for SET, such as their inability to address the origin of functional behaviors or the emergence of functional diversity in biological systems, are not satisfactorily answered.

The Weak Etiology theories (WET), championed by Buller (1998) and Kitcher (1993), claim that the function of a trait is the contribution of this trait to the natural selection through time of the kind of organism which has this trait. The key strategy of this approach is to shift the focus from the evolutionary history of the trait to the evolutionary history of the system. The criterion for determining if a trait is functional is not whether this trait has contributed by a given effect to its own preservation, but to the fitness of systems of the type to which this trait belongs, regardless of whether any other potential biological alternatives have been removed in the process of natural selection.

Thus, WET is less restrictive than SET. All these proposals appeal to the evolutionary history of the function bearer in order to ground functions, but WET does not specify that a trait has been selected *because of* its function. This allows it to account for many cases of functional traits that the other evolutionary theories let aside.

Within this evolutionary framework, there is also a non-historical evolutive approach: the Propensity View (Bigelow and Pargetter 1987; Canfield 1964; Ruse 1971). This approach identifies functions with current causal contributions of components to the life chances (or fitness) of the current systems. A function is a trait's effect which causes the trait to be evolutively selected in the future. In Bigelow and Pargetter terms:

> What confers the status of a function is not the sheer fact of survival-due-to-a-character, but rather, survival due to the propensities the character bestows upon the creature. (Bigelow and Pargetter 1987, p. 187)

This is a forward-looking evolutionary approach. A function is a current contribution to the fitness of the organism, which is an evolutionary cause of the existence of future instances of this organism type. According to PV "something has a (biological) function just when it confers a survival enhancing propensity on the creature that possesses it" (Bigelow and Pargetter 1987, p. 188). Thus, a function of a trait depends on the ways in which this trait will behave in future selective regimes: the biological function F of a trait Y in an organism S is the current effect of T which (presumably) will be the cause of the natural selection of S and, consequently, of the future existence of organisms like S with traits like T.

The main problem of the evolutionary perspective is that every evolutionary account is *epiphenomenal*: according to this historical-evolutionary view, functional attributions have no relation to the current contribution of the trait to the system, since they point solely to the selective history of the trait (Christensen and Bickhard 2002; Mossio et al. 2009, p. 821). As Mossio et al. hold:

> [The evolutionary] theories provide an account that is problematically epiphenomenal, in the sense that it maintains that the attribution of a function does not provide information about the 'phenomenon' (the current system) being observed. From the perspective of the SE theories, a function does not describe anything about the current organization of the system being analyzed. (Mossio et al. 2009, p. 821)

This epiphenomenal character is problematic because it is at odds with the fact that functional attributions seem to have a relation – captured by the causal role approaches – to what function bearers currently do, and not only to the causes of their current existence.

Even the forward-looking Propensity View is also epiphenomenical, but in a different sense to the historical theories. In this case, a function ascribed to a token trait does not explain this token trait, but the future existence of other tokens of the same type. Accordingly, the current heart pumping may explain why this trait type will be evolutionarily selected, allowing the existence of offspring with hearts that pump blood in the future, but the current pumping of blood has no explanatory power with respect to the existence of the current heart.

In conclusion, the Evolutionary Approach is able to address the teleological dimension of the concept of function avoiding the problem of the infra-determination of the causal-role theories. However, all these evolutionary approaches ground this teleological dimension appealing to different systems' token features. In trying to answer teleological questions such as "why a trait T exists?" or "Why a trait X will exist in the future?" both historical and propensitivist theories provide a characterization of "causal loop" that appeals to different trait tokens. By affirming that the function of a trait explains the existence of this trait, these explanations actually refer to the "type" and not to the "token" trait. That is why the evolutionary perspective is vulnerable to various criticisms and objections such as its inability to account for the origin or emergence of new functions.

5 Functions as *Causal Roles* and *Causes of Existence*: The Organizational Approach

The recent Organizational Approach (OA) is an integrative proposal that adopts the opposite strategy to the RT. Instead of offering a splitting account considering different systems in order to justify the teleological loop that justifies that a function explains the existence of a trait, the OA aims to provide a unified definition of functions by extending the teleological dimension to the current activity of a trait.

Functional attributions to both past and current traits explain the presence of the very trait in terms of the effects of its contribution to the self-maintenance of the system to which it belongs. Biological beings are self-maintaining systems since they realize a specific kind of causal regime in which the action of a set of parts is a condition for the persistence of the whole organization through time. Thus, the organizational concept of function applies to classes of self-maintaining systems in current or past regimes of self-maintenance, by preserving in both cases its teleological and normative dimensions.

The notion of self-maintenance comes from a theoretical and mathematical framework developed over the past 40 years by an increasingly rich body of scientific literature. In theoretical biology, complex systems theory, and far-from-equilibrium thermodynamics, self-maintenance refers to a specific causal regime,

realized by various kinds of natural systems, by which a given system is able to exert a causal influence on its surroundings in order to preserve the boundary conditions required for its own existence. In its minimal form, this is shown in the so-called "dissipative structures" (Glansdorff and Prigogine 1971; Nicolis and Prigogine 1977), i.e. systems in which a macroscopic ordered pattern (a "structure"), emerging in the presence of a specific flow of energy and matter in far-from-thermodynamic equilibrium boundary conditions, exerts a constraining action on its boundary conditions that contributes to the maintenance of that FFE flow of energy and matter required for its own persistence. In nature, a very broad set of physical and chemical systems, such as Bénard cells, flames, whirlwinds, hurricanes, and oscillatory chemical reactions can be pertinently described as self-maintaining dissipative systems (Chandresekhar 1961; Field et al. 1972; Field and Noyes 1974).

The different formulations proposed, among others, by Schlosser (1998), Collier (2000), Bickhard (2000, 2004), McLaughlin (2001), Christensen and Bickhard (2002), Delancey (2006), Edin (2008) and ourselves (Mossio et al. 2009; Saborido et al. 2011), base the grounding of the functional attributions in this self-maintaining organization of biological systems.

In a self-maintaining organization, functions can be interpreted as specific causal effects of a part or trait, which contribute to generate a complex web of mutual interactions, which, in turn, maintains the whole organization and, consequently, the part itself. Organizational theories argue that there is a causal loop at the basis of biological organizations, based in the processes of self-maintenance. This causal loop allows us to ascribe a function to a specific trait to the extent that, due to that trait's disposition that we label "function", the trait contributes to the maintenance of the biological organization to which it belongs.

Since self-maintenance of living systems is possible only insofar as the adequate boundary far-from-equilibrium conditions are maintained, and since the structure itself contributes to maintaining these conditions, the activity of the system becomes a necessary (even if not sufficient) condition for the system itself. The system has to maintain an appropriate interaction with its surroundings to maintain itself. Organizational approaches, such as the one defended by us, claim that organizational closure constitutes the relevant causal regime in which the teleological and normative dimensions of functions can be adequately naturalized. Therefore, that which a self-maintaining system does is relevant; it makes a difference in itself, since its very existence depends on the effects of its activity. An organizational function is therefore a condition for the existence (self-maintenance) of the function bearer. Moreover, such mutual dependence between existence and activity, which is specific to self-maintaining systems, provides an intrinsic and naturalized criterion to determine what norms the system, and its parts, are supposed to follow.

Elsewhere (Mossio et al. 2009; Saborido et al. 2011), I have defended my own version of OA. According to this account, the specific regime of self-maintenance that grounds functionality is what we call "organizational closure". This concept is of increasing importance in theoretical biology and philosophy of biology (see Chandler and Van De Vijver 2000 and Mossio and Moreno 2010) and it is a key

notion to understand the specific kind of organization of living beings. In Mossio et al.'s words:

> Biological systems generate a network of structures, exerting mutual constraining actions on their boundary conditions, such that the whole organization of constraints realizes collective self-maintenance. In biological systems, constraints are not able to achieve self-maintenance individually or locally: each of them exists insofar as it contributes to maintain the whole organization of constraints that, in turn, maintains (at least some of) its own boundary conditions. Such mutual dependence between a set of constraints is what we call closure, the causal regime that, we claim, is paradigmatically at work in biological systems. (Mossio et al. 2013)

The most typical example of organizationally closed systems are biological systems, and the intimate association between complexity and integration at work in an organizationally closed organization is the relevant ground of functional discourse in Biology. The interplay between a set of mutually dependent structures acting as constraints, each of which makes a specific and distinct contribution, realizes self-maintenance by maintaining the boundary conditions at which the whole organization, as well as its various structures, can exist. In organizational closure, each process or part is, to use Bickhard's terms, *dynamically presupposed* by the other processes and parts in the overall self-maintenance of the system, such that the whole network must work in a specific and adequate way, for otherwise, because of its FFE nature, the system would disintegrate.

In this framework, functional ascriptions are explanatory because they refer to the net of mutually dependent constraints that contribute to the maintenance of an organization upon whose maintenance their own existence depends.

According to our organizational definition (Mossio et al. 2009; Saborido et al. 2011), a trait is functional if, and only if, it is subject to organizational self-maintenance in a system. This definition implies the fulfillment of three different conditions.

A trait T has (or serves) a function F if and only if:

C1. T contributes to the maintenance of the organization O of S;
C2. T is maintained under some constraints exerted by O;
C3. S realizes organizational closure.

> Accordingly, the heart has the function of pumping blood since pumping blood contributes to the maintenance of the organism by allowing blood to circulate, which in turn enables the transport of nutrients to and waste away from cells, the stabilization of body temperature and pH, and so on. At the same time, the heart is produced and maintained by the organism, whose overall integrity is required for the ongoing existence of the heart itself. Lastly, the organism is organizationally differentiated, since it produces numerous other structures contributing in different ways to the maintenance of the system. (Mossio et al. 2009, p. 828)

In sum, living systems are characterized by the possession of different parts, produced within and by the system, that contribute differently to the maintenance of the organization and thus, of themselves. In this way, teleological and normative functional attributions to each biological trait participating in the organizational closure are justified and grounded.

6 Conclusions: A New Taxonomy of Functions

In a paper by Walsh and Ariew published in 1996, they developed a "taxonomy of functions". There, Walsh and Ariew explain the different formulations of the different theoretical approaches as well as their mutual relations within the philosophical debate on functional explanations of that moment (Walsh and Ariew 1996). This taxonomy offered a panoramic review of the philosophical discussion and Walsh and Ariew used it to introduce their own proposal, the relational theory. Walsh and Ariew claimed that every function is a C-Function, i.e., a function according to Cummins definition, and distinguished between "Causal Role" and "Evolutionary" theories. Evolutionary Functions (E-Functions) are teleological and they are divided in Current (or Propensitivist) and Historical views.

I think that the taxonomy of Walsh and Ariew is still essentially right and does a good job at showing the state of the theories on functions at that time. However, the philosophy of biology has changed significantly in recent years and new approaches and theories have emerged in the philosophical debate on functions. Thus, a different contemporary taxonomy of functions should be formulated in order to account for the current state of the art, by taking into account the new approaches in the etiological and dispositional views and integrating this new organizational perspective.

In this paper, I have introduced an update of this description, emphasizing the novelties of the last years. These new approaches have changed the "geography" of the philosophical debate on the concept of biological function.

According to the current state of the philosophical discussion, I propose the following taxonomy of functions (Fig. 1):

This taxonomy shows that all functional theories are dispositional. And there are two main kinds of theories of functions: causal-role theories (which include SA, GCA and the new OA) and evolutionary theories (where we find the historical approaches of SET and WET and the forward-looking perspective of PV). This taxonomy highlights the fact that both evolutionary approaches and the organizational approach are teleological. Moreover, OA, SET and WET are etiologic. Let me clarify this in more detail.

As I have explained, in a general sense every function is a disposition and there are two main ways to ground the kind of disposition that can be interpreted as a function. On the one hand, the different causal-role approaches base the theoretical grounding of functions on the disposition of a specific trait to contribute to achieve a concrete systemic goal or capacity. On the other hand, evolutive functions are also dispositions. According to the evolutionary account, a biological trait's effect is a function if it entails a disposition to contribute, either in past instances or in current organisms, to the selection of the trait via natural selection (as defended by the SET) or to the fitness of the past (WET) or present (PV) organisms.

Besides this dispositional character, the concept of biological function is interpreted by some theories as a teleological one. The Evolutionary Approach claims that a function has an explanatory role, in evolutionary terms, with regards to

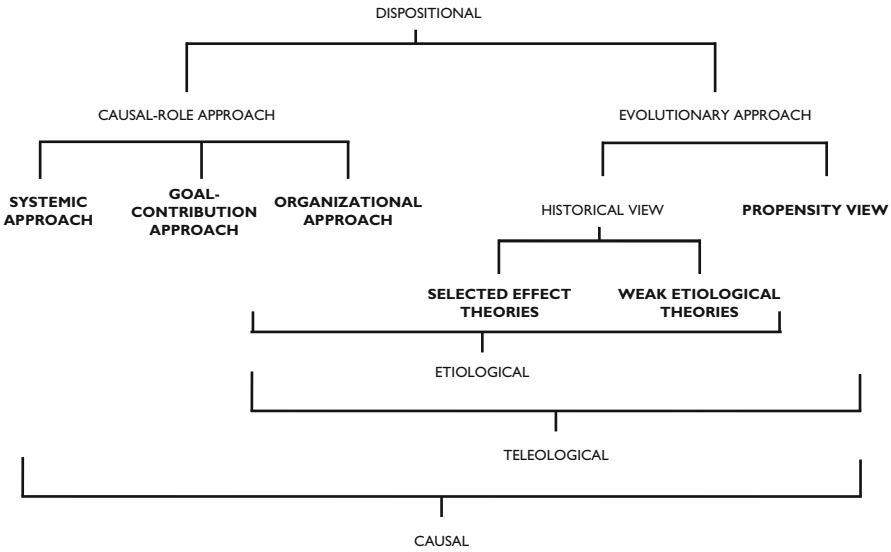

Fig. 1 A new taxonomy of functions, which includes the new Organizational Approach

the present or future existence of the function bearer. And the same teleological character is present in the case of the Organizational Approach, a systemic approach that grounds functional adscriptions in a non-historical and, consequently, non-epiphenomenal causal-loop.

As noted, the explanatory strategy of historical evolutionary theories (SET and WET) is etiological. And, from a non-historical perspective, the same applies to OA. SA and GCA approaches are neither teleological nor etiological. And PV is evolutionary and teleological but, by having a forward-looking strategy, it is not etiological.

In conclusion, this paper has tried to offer a panoramic view of the current state of the debate on biological functions, emphasizing the relevance and novelty of the new organizational strategy. In a discussion that can be mainly understood as a confrontation between two principal stances, the OA perspective introduces itself as a half –way route capable of gathering the best of each of these approaches.

Thus, on the one hand, there are the classic systemic approaches (SA and GCA) that are unable to adequately ground the normative dimension of functional explanations and underdetermine the conditions for the adscription of biological functions because they reject the teleological character of functions. On the other hand, there is the evolutionary approach, which is, in the historical or current alternatives, problematically epiphenomenal.

The organizational approach aims to combine the etiologic and systemic approaches in a teleological and non-epiphenomenal definition of biological functions. As is stated in the taxonomy described in this paper, the new OA is a causal-role, teleological and etiologic approach. This way, it is able to provide clear

criteria for the ascription of functions. However, unlike the historical evolutionary approaches, OA avoids the problem of epiphenomenalism. OA appeals to a causal-loop present in the current living systems, focusing therefore on the biological properties of current organizations.[6]

Therefore, according to organizational formulations, a function has an explanatory role with regards to the very existence of the functional trait. The reasons for the existence of the functional traits are naturalistically grounded in the organizational features of biological systems, interpreted as self-organizing and self-maintaining entities.

In this sense, a trait's effect that contributes to the self-maintenance of the organization is a normative function that is *at the same time* a biological trait's disposition and a cause of existence (in organizational terms) of such trait. Thus, OA reflects the Kantian interpretation of living beings as "natural purposes" by arguing that, indeed, biological functions are both cause and effect of a biological trait. The integration of the concepts of dispositional effect and cause of existence in this new approach opens the way for a naturalized grounding of the notions of teleology and biological normativity and has some important theoretical implications that seem to go beyond the realm of the philosophy of biology.

Acknowledgement The author wish to thank Alba Amilburu, Susana Monsó, Christian Sachse and Marcel Weber for valuable feedback on earlier versions of this paper. The work was funded by UNED ("Proyectos de Investigación propia 2013").

References

Achinstein, P. 1977. Function statements. *Philosophy of Science* 44: 341–367.

Adams, F.R. 1979. A goal-state theory of function attributions. *Canadian Journal of Philosophy* 9: 493–518.

Allen, C., M. Bekoff, and G.V. Lauder (eds.). 1998. *Nature's purposes*. Cambridge, MA: MIT Press.

Ariew, A.R., R. Cummins, and M. Perlman (eds.). 2002. *Functions*. Oxford: Oxford University Press.

Bedau, M.A. 1992. Goal-directed systems and the good. *The Monist* 75: 34–49.

Bickhard, M.H. 2000. Autonomy, function, and representation. *Communication and Cognition Artificial Intelligence* 17(3–4): 111–131.

Bickhard, M.H. 2004. Process and emergence: normative function and representation. *Axiomathes: An International Journal in Ontology and Cognitive Systems* 14: 121–155.

Bigelow, J., and R. Pargetter. 1987. Functions. *Journal of Philosophy* 84: 181–196.

Bird, A. 2007. *Nature's metaphysics: Laws and properties*. Oxford: Oxford University Press.

Boorse, C. 1976. Wright on functions. *Philosophical Review* 85: 70–86.

Boorse, C. 2002. A rebuttal on functions. In *Functions*, ed. A. Ariew, R. Cummins, and M. Perlman, 63–112. Oxford: Oxford University Press.

[6]See Saborido et al. (2011) for an organizational answer to the challenge of the ascriptions of biological functions to cross-generation traits.

Buller, D.J. 1998. Etiological theories of function: A geographical survey. *Biology and Philosophy* 13: 505–527.

Buller, D.J. (ed.). 1999a. *Function, selection, and design*. Albany: SUNY Press.

Buller, D.J. 1999b. Natural teleology. In *Function, selection, and design*, ed. D.J. Buller, 1–28. Albany: SUNY Press.

Canfield, J. 1964. Teleological explanation in biology. *The British Journal for the Philosophy of Science* 14: 285–295.

Chandler, J.L.R., and G. Van De Vijver (eds.). 2000. *Closure: emergent organizations and their dynamics*, vol. 901. New York: Annals of the New York Academy of Science.

Chandresekhar, S. 1961. *Hydrodynamic and hydromagnetic stability*. Oxford: Clarendon.

Christensen, W.D., and M.H. Bickhard. 2002. The process dynamics of normative function. *The Monist* 85(1): 3–28.

Collier, J. 2000. Autonomy and process closure as the basis for functionality. *Annals of the New York Academy of Sciences* 901: 280–291.

Craver, C.F. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–765.

Cummins, R. 2002. Neo-teleology. In *Functions*, ed. A. Ariew, R. Cummins, and M. Perlman, 157–172. Oxford: Oxford University Press.

Davies, P.S. 2001. *Norms of nature. Naturalism and the nature of functions*. Cambridge, MA: The MIT Press.

Davies, P.S. 2009. Conceptual conservatism: The case of normative functions. In *Functions in biological and artificial worlds. Comparative philosophical perspectives*, ed. U. Krohs and P. Kroes, 127–146. Cambridge, MA: The MIT Press.

Delancey, C. 2006. Ontology and teleofunctions: A defense and revision of the systematic account of teleological explanation. *Synthese* 150: 69–98.

Edin, B. 2008. Assigning biological functions: Making sense of causal chains. *Synthese* 161: 203–218.

Field, R.J., and R.M. Noyes. 1974. Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction. *Journal of Chemical Physics* 60: 1877–1884.

Field, R.J., E. Körös, and R.M. Noyes. 1972. Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in the bromate-cerium-malonic acid system. *Journal of the American Chemical Society* 94: 8649–8664.

Glansdorff, P., and I. Prigogine. 1971. *Thermodynamics of structure, stability and fluctuations*. London: Wiley.

Godfrey-Smith, P. 1994. A modern history theory of functions. *Noûs* 28: 344–362.

Gould, S.J., and E.S. Vrba. 1982. Exaptation: a missing term in the science of form. *Paleobiology* 8: 4–15.

Griffiths, P.E. 1993. Functional analysis and proper functions. *The British Journal for the Philosophy of Science* 44: 409–422; also in: Allen, C., M. Bekoff, and G.V. Lauder (eds.). 1998. *Nature's purposes*, 435–452. Cambridge, MA: The MIT Press.

Hardcastle, V.G. 2002. On the normativity of functions. In *Functions*, ed. A. Ariew, R. Cummins, and M. Perlman, 144–156. Oxford: Oxford University Press.

Kitcher, P. 1993. Function and design. *Midwest Studies in Philosophy* 18: 379–397.

Krohs, U., and P. Kroes (eds.). 2009. *Functions in biological and artificial worlds. Comparative philosophical perspectives*. Cambridge, MA: The MIT Press.

McLaughlin, P. 2001. *What functions explain. Functional explanation and self-reproducing systems*. Cambridge: Cambridge University Press.

McLaughlin, P. 2009. Functions and norms. In *Functions in biological and artificial worlds. Comparative philosophical perspectives*, ed. U. Krohs and P. Kroes, 93–102. Cambridge, MA: The MIT Press.

Melander, P. 1997. *Analyzing functions. An essay on a fundamental notion in biology*. Stockholm: Almkvist & Wiksell International.

Millikan, R.G. 1989. In defense of proper functions. *Philosophy of Science* 56: 288–302.

Millikan, R.G. 1993. Propensities, exaptations, and the brain. In *White queen psychology and other essays for Alice*, ed. R.G. Millikan, 31–50. Cambridge, MA: The MIT Press.
Millikan, R.G. 2002. Biofunctions: two paradigms. In *Functions*, ed. A. Ariew, R. Cummins, and M. Perlman, 113–143. Oxford: Oxford University Press.
Mossio, M., and A. Moreno. 2010. Organisational closure in biological organisms. *History and Philosophy of the Life Sciences* 32: 269–288.
Mossio, M., C. Saborido, and A. Moreno. 2009. An organizational account for biological functions. *The British Journal for the Philosophy of Science* 60(4): 813–841.
Mossio, M., L. Bich, and A. Moreno. 2013. Emergence, closure and inter-level causation in biological systems. *Synthese* 78(2): 153–178.
Neander, K. 1991. Function as selected effects: The conceptual analyst's defense. *Philosophy of Science* 58: 168–184.
Nicolis, G., and I. Prigogine. 1977. *Self-organisation in non-equilibrium systems: From dissipative structures to order through fluctuation*. New York: Wiley.
Popper, K. 1959. *The logic of scientific discovery*. London: Hutchinson & Co.
Price, C. 1995. Functional explanations and natural norms. *Ratio (New Series)* 7: 143–160.
Price, C. 2001. *Functions in mind: A theory of intentional content*. Oxford: Oxford University Press.
Rosenblueth, A., N. Wiener, and J. Bigelow. 1943. Behavior, purpose and teleology. *Philosophy of Science* 10: 18–24.
Ruse, M. 1971. Functional statements in biology. *Philosophy of Science* 38: 87–95.
Saborido, C., M. Mossio, and A. Moreno. 2011. Biological organization and cross-generation functions. *The British Journal for the Philosophy of Science* 62(3): 583–606.
Schlosser, G. 1998. Self-re-production and functionality: A systems-theoretical approach to teleological explanation. *Synthese* 116: 303–354.
Schwartz, P.H. 1999. Proper function and recent selection. *Philosophy of Science* 66: 210–222.
Shoemaker, S. 1980. Causality and properties. In *Time and cause: Essays presented to Richard Taylor*, ed. P. van Inwagen, 109–135. Dordrecht: Reidel.
Sommerhoff, G. 1950. *Analytical biology*. Oxford: Oxford University Press.
Walsh, D.M. 2008. Teleology. In *The Oxford handbook of philosophy of biology*, ed. M. Ruse, 113–137. Oxford: Oxford University Press.
Walsh, D.M., and A. Ariew. 1996. A taxonomy of functions. *Canadian Journal of Philosophy* 26: 493–514.
Wouters, A.G. 2005. The function debate in philosophy. *Acta Biotheoretica* 53(2): 123–151.
Wright, L. 1973. Functions. *Philosophical Review* 82: 139–168.

Part III
Philosophy of the Cultural and Social Sciences

How Essentialism Properly Understood Might Reconcile Realism and Social Constructivism

Wolfgang Spohn

1 Introduction

This paper is intended to be about social ontology. There are indeed many specific problems about social ontology, as revealed in the relevant literature, that are most fascinating and that may be independent from foundational ontological issues. However, my fear is that unclarities about ontology in general radiate to social ontology, and therefore I want to start with ontology in general.

But it is not only this fear that drives me. It is also the hunch that no little interest in social ontology derives from a fundamental ontological divide. There is realism claiming that reality is basically mind-independent, and there are various brands of idealism or (social) constructivism united in the claim that reality is basically mind-dependent. Certainly, the first question then is what mind-dependence could mean here. In any case, it may seem that the two opposites meet in social ontology; that is, it may seem that mind-dependence might hold for social ontology and mind-independence otherwise, so that each side is right halfway.

It is also for this reason that I first turn to ontology in general. And I will also end up concluding that realism and social constructivism are both right halfway. However not in the way just envisaged, but rather concerning ontology *tout court*. This will still have interesting implications for social ontology, as I will briefly explain at the end of the paper. In the main, though, I will discuss general ontology.

Let me first flesh out a bit the basic opposition. Realism is the NOA, the natural ontological attitude.[1] There are some things, cars, for instance, and other artifacts,

[1] If I may say so, in order to recapture "NOA" from its displaced usage by Fine (1984). However, I won't discuss here the fine distinction between everyday and scientific realism.

W. Spohn (✉)
Department of Philosophy, University of Konstanz, Universitätstr. 10, Konstanz 78457, Germany
e-mail: wolfgang.spohn@uni-konstanz.de

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_17,

which we have made and which thus depend on our minds, in a clear sense. However, most things, stars and stones, trees and bees, numbers and sets are what they are, without us adding anything to them; they would just be so, even if we and our minds didn't exist.

However, we know well enough how idealism creeps in with Berkeley. For Kant, things as appearances, the only objects about which we can know anything, are the product of a synthesis of intuitions. Synthesis is something performed by a subject. Still, it is objective for Kant insofar as it is done by the unique transcendental subject. However, as soon as you give up on that, you end up with each of many empirical subjects performing their own syntheses (in Kantian terms) and thus with the idea of social constructivism that ontology, i.e., which objects exist, depends on our individual and social constructions. Similarly, phenomenologists speak of the constitution of objects as something done by us.[2]

Quine (1960), to mention just one further prominent position, certainly belongs to neither group. However, his realism is shallow, since he only accepts redundant inner-theoretical truth and rejects any trans-theoretical perspective. We speak of the objects of which we speak; and our ontology is determined by our language/theory, which we impose per fiat on foreign linguistic communities, because their ontologies are inscrutable, anyway.

To put the issue in still other terms: There is the common saying that we carve up the world at its joints. But it has two different emphases: there is the realist emphasis that there *is* the world with its joints and we attempt to carve it up there; and there is the constructivist emphasis that *we* carve up the world and the joints are where we carve.

So, how are we to understand or to integrate the natural realism and the constructivist temptation by which many have been seduced? This is the issue I want to address.

2 What Is an Object?

The only way I see for proceeding on this issue is to start right at the beginning, at the fundamental ontological question: What at all is an object? What a question! What may count as an answer is pointed at by the old dictum: no entity without identity. So, we have to look for the identity conditions of objects. In principle, the answer is given by Leibniz' principle, which I prefer to express in a negative way (because of the awkwardness of identity sentences):

a numerically differs from *b* if and only if there is a property which *a* has and *b* lacks.

[2]I am deliberately speaking here in a general way. Any specific reference would stir up a hornets' nest of subtly distinct positions, which can only be misrepresented by short statements. However, Devitt (1991) is still a beautiful representation (and criticism) of various forms of non-realism.

This contains both, the unproblematic indiscernibility of identicals as well as the problematic identity of indiscernibles. The "if and only if" ensures that the principle indeed delivers identity conditions.

The problem with Leibniz' principle is the quantification over properties. Which properties, precisely, are intended here? Does anything expressed by an extensional formula with one free variable count as a property? Then the principle is trivially true. May the free variable occur in intensional contexts of the formula? Then the principle is either trivially false, or one can claim difficulties with interpreting such formulae. Does identity with *a* count as a property? Then, again, the principle is trivially true. Are only non-relational qualitative properties quantified over? Then the principle is trivially false. In any case, the answer seems trivial. Is there a way to turn Leibniz' principle into a true *and* substantial principle?

Yes, I think so. Let us call a property *proper* if it does in no way refer to identity, i.e., if it may expressed by some formula with one free variable and without identity. Thus, proper properties may be relational or non-relational. Restricting Leibniz' principle to proper properties does not yet help, though. Then it is still trivial. For instance, different concrete objects presumably must occupy different places, it seems, at least at some time (and/or in some possible world).

In this context, the crucial notion certainly is that of an essential property: *F* is an *essential property of a* if and only if *a* cannot fail to have *F*, if *a* cannot exist without *F*, i.e., if *a* has *F* in each possible world in which it exists, and thus if every possible object which is not *F* cannot be *a*. It is precisely the ancient or medieval notion of necessity *de re* or metaphysical necessity, rejected by the logical empiricists and also by Quine and recovered by Kripke (1972) and others, which is invoked here. There are improper essential properties like being self-identical, which applies to everything, or like being identical to *a*, which applies only to *a*. And there are proper ones; for instance, being human is essential for me.

It is important that essentiality is a relation: a property is essential *for* an object. One may call a property *essential* simpliciter if each object has or lacks it essentially. Thus, being human is also an essential property simpliciter; nothing is only contingently human. However, when I speak of essential properties in the sequel, I don't speak of the latter; I will rather be sloppily referring to the relational usage.

A crucial observation is that there also are relational essential properties. I am essentially the son of my parents (and thus the grandson of my grandparents, and so on). Someone could be very much like me, even in extreme degrees; if he is not the son of my parents, he could not be me. Thus, I ontologically depend on my parents in the precise sense that I could not exist without them; in any world in which I exist my parents must exist as well, but not vice versa. Likewise, the number 2 is essentially smaller than 3, the root of 4, etc.

If *F* is applicable to *a*, but not an essential property of *a*, then *F* is *contingent* or *accidental* for *a*. This means that *a* may or may not have *F*. Being now here is contingent for me, and having me as a son is an accidental relational property of my parents. The number 2, by contrast, has no contingent properties, at least within its home field of arithmetical properties and relations. It belongs to the realm of necessity.

Let us call the conjunction of all essential properties of *a* the *essence* of *a*. Then, essences can be qualified as proper, non-relational, etc., just as the properties themselves. For instance, the *proper essence* of *a* is the conjunction of all proper essential properties of *a*. Then I think the appropriate, i.e., a true and substantial version of Leibniz' principle is this:

> *a* numerically differs from *b* if and only if their proper essences (including their relational essences) differ, i.e., if there is a proper property which *a* has and *b* lacks essentially.

So, the substantial claim is the form the identity of indiscernibles thereby takes, namely that objects have no haecceities transcending their proper essence. Surely, this version of Leibniz' principle is contested, and there is a long-standing debate about this. I am not starting to defend it and simply presuppose it for the rest of this paper.

However, it looks at least plausible. I am essentially the unique human offspring of that egg of my mother and that sperm of my father. In my view, this includes that I essentially have no monozygotic twin. If that fertilized egg would have divided in two, I would have been none of those twins; and so it would be with other (symmetric) fissions.[3] Thus, among all possible objects, I am thereby uniquely characterized. Similarly, the number 2 essentially has its position in the progression of natural numbers; this characterizes it uniquely and entails all of its other essential properties. I am unsure how seriously one should take alleged counter-examples such as in Leitgeb and Ladyman (2008).

This principle entails that our talk of objects and identity is inseparably bound up with metaphysical necessity; objectual talk is modal talk right from the start. Therefore I think, by the way, that animals don't have our notion of an object. Animals have remarkable ways of identifying objects, and these ways have become ever more reliable and sophisticated in evolutionary history. Still, they can be tricked. We may also be tricked, even with our superior means of identification. The difference is that we have a standard of numerical identity, our distinction between essential and accidental properties, by which we could tell in principle, or from the God's eye view, whether we are tricked. I don't see how animals could do the same, how they could have the same distinction and thus the notion of an object and of identity.

3 The Distinction Between Essential and Accidental Properties

This remark leads me to the next important question: where does this distinction come from? I think, the only good answer leads us right to the core truth of social constructivism: It is *we* who impose this distinction on nature, metaphysical

[3]This view is contested, of course. See, e.g., Lowe (2002, Part I).

necessity is our invention and convention, and since this is bound up with identity and objecthood, it is *we* who constitute objects.

We have to work a little bit in order to understand this properly and to understand in particular why our natural realism is in no way compromised by this answer. Moreover, we shall see later on that this constructivist claim is only half true; there remains space for discovering metaphysical necessity even after our imposition. However, this amendment can only be introduced after elucidating the crude constructivist claim.

Before these elucidations, let me relate this claim to the previous section. I motivated attending the essential/accidental distinction by the essentialist version of Leibniz' principle. However, this distinction is independent from the principle; we may well accept the distinction, as we should in any case, while doubting this version of Leibniz' principle. So, the constructivist claim does not rely on this principle. However, it is only via this principle, via the dependence of ontology or objecthood on this distinction, that this claim entails the stronger claim that *we* construct our ontology in some sense. And it is this stronger claim in which we are ultimately interested.[4]

Let me provide some reasons for the claim above, a positive and a negative one. The negative reason is that I don't see how we could simply discover essential properties in reality. We can find out that something is human, or square; but how do we find out that it is essentially human, or essentially square? This does not seem to be the kind of property to be empirically discovered. However, if nature does not provide the distinction between essential and accidental properties, where does it come from? There remains only one option: Somehow, the distinction is built in into the way we conceive of the world; we add it to the world.

This nicely fits to a general conception of modality. One may take (some) modal facts as brute facts, thus fending off further explanatory demands. However, if one does not want to acquiesce in these mysterious brute facts, then one might either go with Lewis (1986a, pp. ix ff) for Humean supervenience, according to which all modal facts supervene on non-modal facts. Or one might go with Blackburn (1993) for Humean projection, according to which natural modalities like nomic and causal necessity somehow are projections or objectivizations of our subjective propositional attitudes. This would be my preference.[5] And the human origin of the essential/accidental distinction is well in line with the latter conception.

There is also a positive reason. It is that we can simply impose this distinction and thereby constitute new objects. This is no mystery; we, or at least we philosophers, do it all the time. Quine (1960, ch. II) invented rabbit stages. A rabbit stage essentially consists of a certain rabbit; different rabbits, different stages. And it essentially exists at a certain time; it cannot exist earlier or later; and again, different

[4]Thanks to Wlodek Rabinowicz for pointing out this clarification to me.

[5]For a constructive exlication of the metaphor of projection with regard to nomic and causal necessity see Spohn (2012, chs. 12, 14, and 15).

times, different stages. So, Quine constituted novel objects by taking their time of existence to be essential for them. By contrast, the temporal extension of rabbits and other familiar concrete things is contingent.

Similarly, some philosophers (e.g., Davidson 1985) say that an event is individuated by, or essentially is, a certain spatiotemporal region (plus its intrinsic content). While it may be dubious whether these are events in the ordinary sense and whether events should be conceived in a less fragile way (cf. Lewis 1986a, ch. 23), it is very clear how events in this strict sense are constituted and hence what they are. These examples demonstrate my point: obviously, we can create, as it were, objects simply by specifying their supposed proper essence. Well, creation is a causal notion and hence inappropriate. We better speak of constitution or individuation, which is not a causal process.

This observation is crucial for preserving our realist sense. The objects thus constituted are mind-independent; they do not depend on us or on our minds in any causal or counterfactual way. The earth and its continents would exist and be as they are, even if unconstituted, even if there would be nobody around to constitute them. Although we can tell what an object is only if we have constituted it, its being constituted by us is not essential to it. Otherwise, all objects would have to wait for our constitution in order to come into existence – clearly an absurd idea. No, if being constituted by us is a property of objects, it is a contingent one.

We must carefully distinguish here between constitutability and actual constitution. Every possible object must be constitutable or individuable; every possible object is distinguished by its essential properties. This is what our version of Leibniz' principle requires. If there were something the individuating essential properties of which cannot be specified, it would be unclear what *it* is; it would already be illegitimate to speak of *something* here.

Among all these constitutable possible objects there are some actual objects, i.e., those existing in, or inhabiting, the actual world. However, even most of the actual objects remain unconstituted. There are rabbit stages, since there are rabbits. However, even though rabbits and rabbit stages exist for many millions of years and even though we talk of rabbits for thousands of years, it was only Quine who had the crazy idea to constitute rabbit stages and to talk of them. That is, if we think or speak of objects, they first have to be constituted or individuated; only then it is determinate what we think and speak of.

So, our ontology, what we think and speak about, depends on what we happen to constitute. However, what actually exists by far exceeds our ontology in this sense; it comprises also all of the actual, constitutable, but unconstituted objects. What actually exists depends only on the actual world; and it is the same for communities with diverging ontologies. There may be difficulties in mutually translating languages with diverging ontological schemes; but insofar the ontologies consist of actual existences, both are right. Such communities live in their own world only in a metaphorical sense; it is only their *mental* worlds that differ. And finally, what actually exists is only a tiny part of what possibly exists, of the class of all possible, constitutable objects.

In a way, all of this reduces to a platitude: we think and speak only of the objects of which we happen to think and speak; of course, this depends on us. And then there also are many objects of which we don't think and speak. What I have added to this platitude is merely that thinking and speaking of an object presupposes constituting or individuating it; and that this is something we have to do as well.

I just said that there are many constitutable, though unconstituted objects. What are the rules of constitutability? There is no unconstrained liberty. We certainly cannot take, as Meinong roughly did, any consistent or even inconsistent set of properties and declare the existence of a possible object having precisely that set as its essence (cf. Parsons 1980 for a formal account of such views). What is more plausible is that for any co-instantiable set of properties, i.e., for which there is a possible object having them, there is a further object having those properties essentially. An elaborate theory of essentialism and of possible objects would have to specify these rules of constitutability; to my knowledge they are (much too) little investigated. However, this is not our present task.

Actual constitution seems to be a lot of work; after all, we think and speak of very many objects. Of course, it is not. It is not individual work. It is even not contemporary social work, although we may change and enrich our ontology here and there. Mainly, we inherit our ontology from our ancestors by growing into their language and its ontological scheme. However, this should not blind us for the fact that our ontology, the kinds of objects we constitute, is part of our linguistic conventions. Even if we take over the conventions of our ancestors, they remain conventions. Therefore I like to speak of *essentiality conventions* which govern our ontology, our constitution of objects.[6]

Conventions: this sounds so arbitrary, as if we could constitute any ontology we like. Yes, to a large extent we do; this is what I wanted to convey. However, this is not to exclude that there are silly and useful, good and bad conventions. It would be most important and fascinating to explore the rationality behind our ontological or essentialistic conventions. Why do we have the conventions we have? Why, for instance, are we used to constitute persistent things and not stages? And why do we constitute those persistent things as enduring and not as perduring?[7] And so forth. I am not aware of deep investigations couched in these terms. Maybe good answers are given under different headings; maybe the rationality lies in somehow maximizing contingency and hence, since explanations refer only to contingent facts, in somehow maximizing our explanatory reach. In any case, that's my point, we do not find an answer by staring at nature and searching for essences there. We rather must look at ourselves and study our ontological policies.

[6]I take this term from Haas-Spohn (1995, sect. 3.5), where it is introduced and discussed in detail.

[7]The distinction of perdurance vs. endurance of persisting objects is due to Lewis (1986b, pp. 99 and 202ff). The presupposition of my question, that we have an ontology of enduring objects, is a big claim contra Lewis, which I am not going to defend here.

4 Putnam's Insight

Matters are still more complicated. So far, I have contended that we declare which of the properties are essential and which are accidental for objects and that we thereby constitute those objects. But this is not quite what we do. Usually, we only say what *kind* of property is essential for an object and leave it open to empirical inquiry which essential property of that kind the object actually has. This then is an inquiry into the essence of that object. In this way, the essentiality conventions only partially fix the essences of objects; within these bounds, the full determination is taken over by nature itself.

For instance, we declare that, if I am human, I am essentially human. But what that is to be human is unknown and open to investigation. Similarly, we say that I have my parents essentially. This leaves the business to you to find out who my parents are (in which you will only succeed by finding out who my grandparents are, and so on; that is, you will never finish the business).

In principle, this point is clear, since Kripke (1972) explained to us that some metaphysical necessities are a posteriori. However, I prefer to call the point Putnam's insight, because Putnam (1975) argued in a particularly forceful way that a natural kind term essentially applies to objects which stand in an unknown theoretical equality relation to supposed paradigms of that natural kind. For instance, water is what stands in the same-liquid relation to most of our water paradigms; and both is up to empirical and theoretical inquiry, the same-liquid relation and the actual nature of our water paradigms. (*Some* of our water paradigms may turn out not to be water; but there is no standard of comparison on the basis of which it could turn out that most of our water paradigms are not water.) And Putnam (1975, pp. 235ff – his example is "gold") made also very clear that it is our convention to treat terms like "water" as natural kind terms. We could also use "water" as a term essentially applying to anything that has the same superficial characteristics as our paradigms, such as being fluid, colorless, and tasteless. But this would be a different usage. Thus, the convention is to use "water" as a natural kind term, and the precise nature of the natural kind of water is up to discovery.

This allows for the possibility that we do not find any underlying nature. Any natural kind term comes along with a hierarchy of fallback positions governing our responses to unexpected discoveries. If we find only chaos underneath the surface, we might even end up with taking the essence of water to lie in its superficial characteristics; but this would then be the result of investigations, not a conventional ruling right from the start.

These remarks extend to objects. If I am essentially human and if being human is a natural kind, then there is something to find out about my nature. Moreover, if my origin, i.e., my parents are essential to me, this also fixes only a kind of relational property essential to me; and it still leaves the task of finding out who my parents are.

So, Putnam's insight leaves the fact untouched that our usage is governed by essentiality conventions, and this fact is quite explicit in Putnam's work. Emphasizing the insight might have obscured the fact about conventions. Both points are

important, and this is why I have introduced the insight only after arguing for the human origin of the essential/accidental distinction. Still, the insight shifts, in a way, the weights between realism and constructivism in favor of the former, though only to an extent admitted by the latter. And the point puts the above issue about the rationality of our essentiality conventions into a new light. Apparently, it is often reasonable to delegate the fixation of essences to nature within conventional bounds.

Let me summarize: I argued that the distinction between essential and accidental properties and hence the constitution of objects is due to the essentiality conventions of our linguistic community. There is this much truth in social constructivism. In this sense we construct the world. However, this phrase is dangerous and misleading. Construction must not be given any causal meaning here. The world, at least the natural world, and its objects would exist in the very same way, even if our constructions were different or non-existent. Different constructions would speak about different objects; but this does not mean that the unspoken objects do not exist and are not what they are.

The situation is nicely highlighted by the catch question attributed to Abraham Lincoln: how many legs would a monkey have, if we would also call its tail a leg? The right answer is, of course: still 4, not 5. We don't change the world by speaking differently about it. So, despite the social constitution of objects we may stick to our natural realism – all the more as the essence of objects very often is as it is and waiting to be discovered, within the bounds established by our conventions.

5 Consequences for Social Ontology

What does all of this entail for social ontology? The negative conclusion is that social ontology does not provide the special arena in which realism and social constructivism would meet, as I have envisaged in the introduction of this paper. They meet in the general arena in the way indicated.

The positive conclusion is that the general ontological observations apply to social ontology as well. However, this is not to say that social ontology would not have its peculiarities. On the contrary, there are at least two striking differences.

The first difference is that the social world is indeed constructed by us in the ordinary sense. All the objects belonging to it are causally and indeed ontologically dependent on us; they would and could not exist as what they are without us. And they are many: all the artifacts, houses, furniture, clothes, cars, books, banknotes, etc.; our environment is overcrowded by artifacts. An artifact belongs to its kind essentially, like an animal or a plant it has its origin essentially, and thus it has a unique essence. (Since we made the artifacts, we more easily slip into the quandaries of fission, fusion, gradual substitution (as in Theseus' ship), etc. However, they pose problems for everyone, not only for essentialism.)

In principle, the same applies to more abstract social objects, political institutions, nations, social formations, religions, economic organizations, etc. In those areas we find many examples where conceptualizations not only represent, but

indeed make the world, as the social constructivist claims. However, they make the world not in the sense of Goodman (1978), which he extends from the cultural to the natural world and which I find obscure, but rather in the sense of Searle (2010), which I do not find obscure and which basically seems to me to be the ordinary causal sense.[8] These effects may even reach deeply into individual psychology. We may well grant that the mental states and attitudes, even the feelings we actually have are deeply imprinted by how we conventionally conceptualize them. And this is definitely responsible for a lot of foreignness across times, societies, and cultures. In any case, in all these areas there is a lot of our own making.

The second difference we find in social ontology lies in the kind of essential properties. I mentioned that the origin of an artifact is part of its essence; this is no peculiarity. However, we must also say to which kind it essentially belongs; otherwise we don't know which object came into existence at its origin. Here we find a difference; in nature we usually constitute natural kinds, whereas in culture we very often constitute functional kinds. At least this applies to all the kinds of artifacts I have mentioned above.

And it applies to more abstract social entities like money, property, taxes, economic and political institutions and offices, social roles, etc. Let me quote from Weidmann (2012), from the current president of the German Federal Reserve; he says: "Money is defined by its functions. ... Money is a social convention." Searle (2010, ch. 5) says that all those entities derive their existence from our status function declarations and thus from our declarative speech acts. In any case, they have those functions essentially.

This entails that the essences of the objects of our social world are usually not hidden and unknown. Well, this is not quite true; the origin of a particular artifact is often unknown and of no further interest. But it is true of the kinds. Their function is common knowledge; hence we know their essences and thus the kinds themselves. There is no hidden nature of chairs or cars or checks or chancellors.

We may describe this point in a different way. In Spohn (2012, sect. 16.4) I defended the view that an individual person is conscious of precisely those facts that are ipso facto known to her, such as her being in pain, her presently thinking of her son, her believing that Berlin is the capital of Germany, her desiring to make vacations, etc. This characterization allows to extend the notion of consciousness to collective subjects. That is, in precisely this sense, one can say that the social consciousness of a community consists in its common knowledge, because it is precisely common knowledge that is known to be common knowledge. In this sense, one can also say that social ontology is part of social consciousness.

However, this applies only to objects and entities in our own community where we may assume common knowledge of them. In principle, though, what I have called Putnam's insight is relevant also in the social realm. If we visit foreign cultures, we clearly find objects that apparently have some function, though we don't know which; and the most evasive of those objects are linguistic signs. In this

[8] See also Devitt (1991, sect. 13.5). I entirely agree with his criticism of Goodman.

case, the foreigners could show and try to tell us the function; this might include teaching us their language. Then our ignorance is relieved. However, matters are not so simple, for instance, when we find strange things in the tombs of our ancestors, where nobody can give us any explanations. And matters are still harder with more abstract social entities like roles and institutions. What they might have been in illiterate foreign cultures is almost impossible to find out, and even with literate societies it is often difficult, since their signs and languages are social entities themselves and hard to access.

Let it suffice with these remarks on social ontology. They are neither systematic nor particularly revealing. Their only point was to briefly indicate how social ontology falls under general ontology in its specific ways. The main point I wanted to make is how even general ontology is socially determined, as social constructivists might have it, though without thereby undermining our natural realistic attitude in any way.

References

Blackburn, S. 1993. *Essays in quasi-realism*. Oxford: Oxford University Press.

Davidson, D. 1985. Reply to Quine on events. In *Actions and events: Perspectives on the philosophy of Donald Davidson*, ed. E. LePore and B. McLaughlin, 172–176. Oxford: Blackwell.

Devitt, M. 1991. *Realism and truth*, 2nd ed. Oxford: Blackwell.

Fine, A. 1984. The natural ontological attitude. In *The philosophy of science*, ed. R. Boyd et al. 261–278. Cambridge, MA: The MIT Press.

Goodman, N. 1978. *Ways of worldmaking*. Hassocks: Harvester Press.

Haas-Spohn, U. 1995. *Versteckte Indexikalität und subjektive Bedeutung*. Berlin: Akademie-Verlag.

Kripke, S.A. 1972. Naming and necessity. In *Semantics of natural language*, ed. D. Davidson and G. Harman, 253–355 and 763–769 Dordrecht: Reidel; ext. ed.: Oxford: Blackwell 1980.

Leitgeb, H., and J. Ladyman. 2008. Criteria of identity and structuralist ontology. *Philosophia Mathematica* 16: 388–396.

Lewis, D. 1986a. *Philosophical papers*, vol. II. Oxford: Oxford University Press.

Lewis, D. 1986b. *On the plurality of worlds*. Oxford: Blackwell.

Lowe, E.J. 2002. *A survey of metaphysics*. Oxford: Oxford University Press.

Parsons, T. 1980. *Nonexistent objects*. New Haven: Yale University Press.

Putnam, H. 1975. The meaning of meaning. In *Philosophical papers, vol. 2: Mind, language and reality*, ed. H. Putnam, 215–271. Cambridge: Cambridge University Press.

Quine, W.V.O. 1960. *Word and object*. Cambridge, MA: The MIT Press.

Searle, J.R. 2010. *Making the social world. The structure of human civilization*. Oxford: Oxford University Press.

Spohn, W. 2012. *The laws of belief. Ranking theory and its philosophical applications*. Oxford: Oxford University Press.

Weidmann, J. 2012. Money creation and responsibility. Speech at the 18th Colloquium of the Institute for Bank-Historical Research in Frankfurt on 18 Sept 2012.

Social Construction – By Whom?

Matti Sintonen

1 Introduction

There is no canonical definition of social constructionism, but let us take a characterization to start with. Vivien Burr (1995) has summed up some of the common threads of social constructionism roughly as follows. Social constructivists will not take received wisdom for granted, for a number of reasons, some cognitive so as to say, others having to do with the uses of knowledge. First, our observations about the world are (and should) not to be taken at face value since they are socially conditioned – many entrenched beliefs and especially unarticulated attitudes behind these beliefs are or could be false, skewed, and in any case ephemeral. Secondly, our conceptions of the world and society are a result of our history and culture and hence relative to these. Therefore there are no timeless and acultural truths. Third, knowledge arises in a social process and hence is a result of human interaction – therefore it is not a picture or mirror image of nature. Fourth, what passes as knowledge at any given time, or how concepts (or things) are explicitly or implicitly categorized and defined, often has grave consequences for the people categorized and defined. Therefore knowledge is (always, or often) not (just) description and explanation of brute facts has a political or power aspect to it.

There is no shortage of representatives of social constructivists – and of course one can be a social constructionists to different degrees (by not taking it as given that one should subscribe to all of these four features, or by interpreting them differently). My focus in this paper is not on the philosophy of mind and on social science. It is now widely recognized that minds are both causally and conceptually social – in that they could neither exist nor be perceived except in relation to other minds (see Searle 1995, Pettit 1996). Rather, I shall talk about the more sweeping

M. Sintonen (✉)
Department of Philosophy, University of Helsinki, Unioninkatu 40A, Helsinki 00014, Finland
e-mail: matti.sintonen@helsinki.fi

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_18,

claim that all scientific knowledge is socially constructed and could neither exist nor be conceived except in relation to a community of inquirers.

Agreed that knowledge, scientific knowledge included, is socially constructed, what does that mean? Ian Hacking, famously, asked: social construction – of what? What sort of entities have been or could be candidates for the value of the variable X in "Social construction of X"? And: what is the point of claiming that X is socially constructed? It turned out that the variable has ranged over a variety of entities, from the world ("nature" or "reality") to scientists (and people's) beliefs about the world. As regards the point of making these claims Hacking's major finding was in line with Burr's list of features: a social constructionists question received wisdom and accepted definitions and claim that what passes as self-evident is far from it. And of course: that the socially constructed beliefs and attitudes are often bad things.

But there is another question about social construction that need to be addressed, if only because it has undergone a series of upheavals. Given that knowledge is socially construed we can ask: Socially construed – but by whom?

2 Social Construction – By Gentlemen

Early precursors of social constructionism vis-à-vis science include Ludwik Fleck et al. (1979) on *thought collectives* (and syphilis), Michel Foucault (1989) on *epistemes* (and The Order of Things), and Thomas Kuhn on paradigms (and the Copernican Revolution), to mention just a few. Berger and Luckmann (1967) claimed that reality (and not just social reality) is socially construed, a result of social interaction and negotiation. Karin Knorr-Cetina (1983) took it that laboratory scientists are not putting questions to nature (as Bacon and later Kant suggested) since "nowhere in the laboratory do we find the 'nature' or 'reality' which is so crucial to the descriptive interpretation of inquiry". Harry Collins (1983) in turn argued forcibly that as determinants of the outcome (knowledge) reason and logic fall behind social patterns: "Rationality (whatever that means) must play little part in explaining how the world comes to appear as it does".

But the roots of constructionism go well beyond that. In a sense Francis Bacon elaborated on this in his proposal for a division of labour in the House of Solomon. Bacon also noted that man more readily believes what pleases his mind, as well as that knowledge is corrupted by the varieties of idols – not least by the idols of the tribe and of the market. Although he conceived inquiry as a process in which nature is forced to reveal her secrets through series of interrogatories he did not think that the procedure is simple or mechanical. Indeed, nature was there to be found and interrogated, but the onus of interpretation always was with the interrogator. Although Bacon seemed to think that knowledge has value as such he maintained that the pursuit of knowledge merely for the purpose of satisfying intellectual curiosity was one-sided if not outright perverse. Knowledge was to be

harnessed to serve common good (represented by the crown, or more exactly James I). Finally, Bacon steered between the two extremes of empiricism (the Empirics) and rationalism (the Reasoners), although his example of social construction comes from social insects. He made it clear that both views would be one-sided: "the true business of philosophy" is between that of the ant and the spider, viz. of the bee: "for it neither relies solely or chiefly on the powers of mind, nor does it take the matter which it gathers from natural history and mechanical experiments and lay it up in the memory whole, as it finds it, but lays it up in the understanding altered and digested" (Bacon 1878, Book One, Aphorism XCV).

Bacon's vision of science summed up the place of science in society. He also influenced the creation of scientific institutions and institutionalised science, the Royal Society in particular, and was held in high esteem for centuries to come. I suggest that it was one link in the series of imaginaries (indeed, sociotechnological imaginaries) – collective visions of what knowledge and science are and how they are to shape the future. Knowledge is power and nature must be understood before it could be commanded. Advancement of knowledge is not to be conceived as contemplation of true propositions but it should rather be advancement of common good. So influential was Bacon's vision that it was turned into a contract between natural philosophy (science) and society.

Stephen Shapin and Simon Schaffer (1985) have argued, referring to the debate about the credentials of the new experimental philosophy advocated by Robert Boyle that the way facts were established was not through reason (alone) but *via* a collective procedure that took shape in a particular context. That Boyle's experimental philosophy (and not Hobbes' traditional insistence on philosophical certainty) got the upper hand had at least as much to do with fertile social and political ground also *outside* the community of natural philosophers than with the power of argument within their narrow circles.

But again, given that knowledge was socially construed at least in the minimum sense that it was a result of a consorted effort and orchestrated division of labour, who were the labourers? Seventeenth century imaginary was not that of a democratic ideal where anyone could propose anything and where disinterested pursuit of truth ultimately prevailed. Rather, as Shapin (1994) has argued in great detail, knowledge-making was a collective enterprise but not everyone's word counted. Only those who were wealthy enough to enjoy independence could be counted to tell the truth. A gentleman's word could be trusted, a gentleman could have no ulterior motive than to be a humble party to the advancement of truth. Not only was a gentleman's word to be trusted, challenging the word of a gentleman was ruled out by social conventions. Seventeenth century scientific culture was therefore a gentlemanly culture with its collective practices of producing experiments and of witnessing the results then led to the established as facts – and these facts where then recorded in the proceedings. Perhaps, then, the early imaginary of natural philosophy or science was this: knowledge is constructed by Gentlemen.

3 Social Construction and the Social Impulse

It is ironic that the Royal Society chose as its motto the credo: *Nullius in verba*, "On Nobody's Word", hence claiming that man should only trust his own senses and reason. Perhaps it should be best seen as part of the ideology, in the literal Baconian sense, of building an imaginary for the new experimental philosophy. It is interesting to contrast this to a more recent imaginary, the way common values and common visions of the place of science in society as it was drawn by Charles Peirce at the end of the nineteenth century. I shall first outline the argument for the beneficial or even crucial role of the autonomous scientific community from a philosophical and then explicitly sociological angle. The two views agree on the claim that science (and knowledge more generally) is social by nature but disagree on detail.

The first view is Charles Peirce's credo of scientific method: science has become so successful because it builds on the assumption that there is a mind-independent reality and because scientific inquiry has particular moral and social aspects to it. The second is Robert Merton then took a step ahead in his view that the success of science is not based on the fact that scientists are a particular creed of morally superior people but rather on the its *particular form of organised* scepticism. On both views scientists are perceived as disinterested pursuers of basic science virtues such as truth (and information), and on both views the crucial concept is that of an autonomous, well-functioning scientific community.

Charles Peirce, a pragmatist with strongly realist inclinations (and hence, a pragmaticist), argued that science differed from everyday belief systems by being systematic and by relying on method. In "The Fixation of Belief" he elaborated on how the method works. Inquiry begins with doubt forced on the inquirer through observations and others' opinions. In order to eliminate the "irritation of doubt" something must be done to restore harmony and to fix belief. The method of science, unlike other methods of fixing belief excels in that it provides a way of settling differences of opinion in an orderly and permanent fashion. This it does for two kinds of reasons. One has to do with the fact that there is "secondness", the realm of Real things that are "entirely independent of our opinion about them". (Peirce, CP, V).

The second virtue of the scientific method might seems at first sight to be at odds with the first one since it reaches into what Peirce calls the moral and social realm, or the realm of human conduct: "The most vital factors in the method of modern science have not been the following of this or that logical prescription – although these have had their value too – but they have been moral factors" (CP, 7.87). The crucial moral factor has been "the genuine love of truth and the conviction that nothing else could long endure". Modern science is committed to truth even in the face of unpalatable consequences, and the genuine man of science always puts truth above convenience and practical utility. The social factor, nevertheless, is all-important: the "next most vital factor of the method of modern science", he wrote, "is that it has been made social". This in turn has a dual nature. First, facts do not confine to one individual only: "what a scientific man recognizes as a fact of science

must be something open to anybody to observe". The other aspect of sociality has to do with "the solidarity of its efforts": "The scientific world is like a colony of insects in that the individual strives to produce that which he himself cannot hope to enjoy".

As social beings we are, Peirce thought, vulnerable to an extremely strong social impulse that guarantees that doubts will arise – and harmony is sought: "Unless we make ourselves hermits, we shall necessarily influence each other's opinions; so that the problem becomes how to fix belief, not in the individual merely, but in the community" (CP, 5. 378). As a result we have a canonical description and a new answer to the question: social construction by whom. The answer is: the autonomous but social community of scientists and scholars. And let us note that this no longer is the community of gentlemen, those who own land, as in the previous imaginary. This is the community of truth lovers. Here the social nature of inquiry is not opposed to thinking along (pragmatically tinted) realist lines. Rather, the scientific community is what guarantees that "truth will out", eventually.

4 The Ethos of Science

While Peirce gave the canonical philosophical account of why science works, he was not a social scientist like Robert K. Merton. There were many parallels in the thought of the two giants, but also one crucial difference. Where Peirce located one of the cornerstones of the success of science in the moral realm, claiming that scientists, unlike representatives of many other professions, are lovers of truth, Merton gave an explicitly social explanation of the success of science. For Peirce the key to the progress of science is the essentially social and self-corrective procedure that is built into the scientific method. Like Peirce Merton was an ontological and epistemological realist, and both critisized radical relativism (and extreme forms of constructionism). There is no way of conducting or starting inquiry but on the premise that the world – nature – comprises an intelligible order. For both men scientific knowledge was socially (and culturally) mediated, and individual scientists always were introduced into some particular community with a distinct thought style.

Peirce and Merton also agreed that truth is a higher aim than short-sighted utility – and hence they can be said to have shared the imaginary of the First Social Contract described below. Peirce wrote that the point of view of utility "is always a narrow point of view" and, as such, not most likely way to the practical utility in the long run but rather something that would hamper the achievement of goals, both cognitive and practical (Peirce, CP, 1.641). Merton held that making science serve the immediate needs of the society would be counterproductive. He writes that "fundamental scientific knowledge is a self-contained good and that, in any case, it will in due course lead to all manner of useful consequences serving varied interests in society" (Merton 1982, p. 214).

However, there is a point where Merton departed company with Peirce. In Merton's view scientific inquiry, just as any other type of social activity, is governed by social norms, here both technical and moral norms governing scientific conduct. These norms are anchored in communally shared values and they give rise and support in individual scientists personality characteristics that comprise what could be called a scientific mind. Where Peirce had maintained that scientists have special moral qualities, the most important of these being the commitment to truth, Merton offers an explicitly social explanation in terms of the communal mechanism in which motivation and control are organized. Science as an institution is set apart from other forms of social activity not just in its goal of expansion of certified knowledge but in the utterly detailed way in which members of the scientific community are rewarded, encouraged, and punished. "The ethos of science", he writes, "is that affectively toned complex of values and norms which is held to be binding on scientists. The norms are expressed in the forms of prescriptions, preferences, permissions and proscriptions. They are legitimized in terms of institutionalised values" (Merton 1982, p. 5). While Peirce's portrayal does make the social impulse and the cooperative mode of conduct crucial to scientific success Merton takes a step further: there is no need to think that scientists are specially endowed with a moral quality that keeps them on the narrow path towards truth. Rather, whenever deviations occur, the built-in community level features, backed by the *Ethos*, both provide an incentive and the mechanisms for punishment.

5 The First Explicit Social Contract

Peirce's idea – that science works best when left in the hands of the autonomy of science and Merton's idea – that there is a parallel between a healthy society (liberal democracy) and a healthy science. Interestingly where Merton had argued that both from a descriptive and a normative point of view a well-functioning society serves as a model for science, Michael Polanyi reversed, in his Republic of Science, the order. As his title suggests, "the community of scientists is organized in a way which resembles certain features of a body politic and works according to economic principles similar to those by which the production of material goods is regulated". He held that that it is in "the free cooperation of independent scientists we shall find a highly simplified model of a free society, which presents in isolation certain basic features of it that are more difficult to identify within the comprehensive functions of a national body". Scientists, whilst freely pursuing their own choice of problems and making their own personal judgment, "are in fact co-operating as members of a closely knit organization" (Polanyi 1962, p. 54). In the decades to come this cornerstone of scientific autonomy became consolidated in a more or less explicit social Contract. The issue was of course about the governance of science. Who is or should be entrusted with the governance of science? What sort of aims should be furthered?

It would of course not be correct to claim that science before World War II was carried out without governance, but at least in the US it was mainly business

conducted at the privately funded research universities. But during World War II the scientific community had shown its loyalty to the government (and the military in particular, after all one outcome was nuclear weapons) thus contributing to US-led allied victory. As a result of this a number of scientists, with Vannevar Bush at the front, began scheming on the dream of a basic scientist: secure and growing public financing of science, with decreasing government control. President Roosevelt then commissioned from Bush a plan for science policy, and the result is what Donald E. Stokes (1994) called a Treatise – that was what the document, *Science: the Endless Frontier,* in effect was.

Fundamental to the Treatise – a Social Contract – was the notion that basic research is a driver of applied inquiry and development, and eventually of technological innovations. As Stokes notes, Bush outlined this notion – the linear model – in two Bacon-inspired aphorisms: first, that basic science is not performed with an eye on practical utility, and secondly, that when not harvested prematurely the results of basic science will foster industrial progress though technology transfer. This was the imaginary for an explicit Social Contract between science and society: politicians should not be trusted with governance of science, the scientific communities should be given a free hand in deciding what to research and how. As a result, the US established NSF. Although its constitution made it clear that [T]he object of the fund shall be the promotion of human welfare through the advancement of science" it nevertheless functioned under the premises that this goal will be achieved through individual scientists and scientific communities that pursue truth and information (and perhaps other cognitive aims) and not welfare or common good directly.

The outcome was a Contract based on an imaginary that was not obvious to all parties. As Harvard historian of science I.B. Cohen, a member of the "secretariat" to Bush report, notes, many scientists feared that government financing endangered autonomy in that it was likely to favour practical projects with predictable results; that the scientific agenda, despite wishful thinking, nevertheless was subject to political interference; that geographical and other considerations might override academic excellence; that federal funding could favour accepted lines of research and not support unorthodox openings; that such a concentration of funding under one or few umbrellas might lead into trouble and lack of funding at hard times; etc. (see Stokes 1994). And indeed the success of NSF was not immediate: ideology or imaginary was strong, but it took years before it was materialized.

6 The New Social Contract

Arguably we now organize our research under a new social contract between science and society. I shall call it the New Social Contract although, given Seventeenth century gentlemanly culture and the rise of scientific societies, what is old and new only makes sense with respect to twentieth century. The reasons and causes for this recontextualization are many, and I can only note some. But the outcome is relatively clear: were the circle of signatories under the First Contract included the

autonomous scientific communities and science funders, most notably in the US the government and the universities, they comprise a whole gamut of other stakeholders from non-governmental organizations and private philanthropists to churches and courts of law. All these signatories have changed our imaginaries concerning the proper aims and even methods of science. It is not clear what the answer to our initial question, "Social Construction – by Whom?" is except perhaps: us all?

Some parts of the First Contract (or treaty) between the two regimes, science and society, still linger on (in the self-understanding of scientists and in the imaginaries of science more generally), but on the whole the Old Contract has been jettisoned or radically altered. One of the drivers of the Second Contract has been the true globalization of science. And as has been noted, this process is far from smooth, nor is it uniform across nation states and continents. As one consequence there is the tension between basic inquiry and applied inquiry (and development and technology, to wit), now recontextualized as academic excellence and relevance. To see how this might have emerged, consider the scientific dream on which the First Contract was based, viz. increased public funding and decreased government regulation. The pressure to strip down has been easy to note. For one, it became increasingly clear that the linear model of science – from basic science to applied science to development and finally technological innovations – was factually inaccurate. Secondly, and relatedly, there was the lingering suspicion that when left on its own science is not just an endless frontier but an endless black hole for money and resources – without guarantees of tangible results.

Recall the terms of the Contract. Scientists are disinterested hunters of cognitive value such as truth and information content, that is, basic science values. But NSF was by its own words, established to promote human welfare. Nor was this just the working in the US. How about OECD? It exists, according to its charter, to promote "the highest sustainable economic growth and employment and a rising standard of living in Member countries" (OECD 1966, p. 19). Nevertheless OECD professed to be a keen supporter of basic research which, by its own definition, "had as its aim the extension of knowledge for its own sake, that of applied research the utilisation of existing knowledge". Ignoring, for now, the outmoded characterisation of the latter, one could easily ask: extension of knowledge for its own sake is a laudable thing, and pouring money to increasingly expensive infrastructure no doubt enhances understanding and the scientists' satisfaction. But then, presumably, the more money you would invest in basic research, the more practical utility you would get? And: why should we, the ordinary citizens represented by politicians and policy makers care about increase in understanding – understanding that is beyond the reach of an ordinary taxpayer? Why should scientists be given open hands on what they investigate, and a bottomless source of money to enjoy their games of maximizing epistemic utility? Given that there are these pressing practical problems shouldn't science be governed by democratic principles, shouldn't it be accountable and responsible?

One of the results of the crumbling First Contract was precisely a new orientation towards directed funding and what has become to be known as strategic funding. According to one definition, "[S]trategic research [is] basic research carried out

with the expectation that it will produce a broad base of knowledge likely to form the background to the solution of recognized current or future practical problems" (Irvine and Martin 1984). Here is how the so-called MASIS report formulates the new type of research: "Strategic research combines relevance (to specific contexts, possibly local) and excellence (the advancement of science as such). The contrast between fundamental (and scientifically excellent) research on the one hand and relevant research on the other hand is not a contrast of principles. It has more to do with institutional division of labour than with the nature of scientific research".

Increasingly often this strategic research is organized around so-called Grand Challenges, usually global problems such as climate change, scarcity of fresh water or ageing. In these cases it is no longer up to the scientists to determine what they want to investigate, unless they have funding from other sources. All these trends indicate that bottom-up scientist-initiated research is on the decrease.

On the face of it these characterizations do not solve the tension between the two principal goals of science and inquiry. The first definition does not specify the time span on which strategic research is to carry fruit – so that the there is but a difference of degree between strategic research and applied inquiry (as it was understood in the First Contract). Nor is there any closer analysis as to how the three types of research, call them pure basic research, strategic basic research and applied research are distinguished (for an early attempt to explain why the distinction is difficult to draw, see Sintonen 1990).

7 Recontextualisation of Science

There have also been other important changes that have led to the second contract – to the extent that it is now customary (in science and technology studies) to speak of the recontextualisation of science: the requirement of relevance has been coupled with new modes of knowledge production.

That science has been in the process of recontextualization can be seen in the blurring of disciplinary boundaries and various forms of interdisciplinary collaboration, in new forms knowledge production (private/public) partnerships, in the triple helix of governance and as the emergence of new agents that attempt to have their voices heard. Here is how the so-called MASIS-report describes the stakeholders in science and research: the Second Contract involves the third sector, media, private companies as well as the general public as stakeholders. Furthermore, there is a new awareness of responsible development, and more attention is given to ELSA aspects of science and as well as to "interactive forms of technology assessment; and experiments with public engagement" (The MASIS Report 2009, pp. 4–5). For example, citizens/users, scientists, local politicians and other stakeholders could be brought to negotiation table concerning water management with the aim of helping local populations adapt to climate change.

I shall conclude with an example that has engaged many philosophers of science recently, viz. the notion of well-ordered science and the case of neglected diseases in

particular (as developed by Philip Kitcher and Julian Reiss, among others) and with advancement in synthetic biology that might respond to the request. That a practice of the sciences is well-ordered amounts to the requirement that inquiry is directed so as to promote the common good, where common good is understood as something that aims "at the goals that would be endorsed in a democratic deliberation among well-informed participants committed to engagement with the needs and aspirations of others" (Quoted from Reiss and Kitcher 2008, see also Flory and Kitcher 2004).

The problem of neglected diseases arises when we ask: how come the lion's share of research and development in pharmaceutics goes to drugs developed for the affluent west (often to diseases created by high living standards)? Why not follow the fair-share principle: common good should be perceived globally and funding should be directed in proportion to the needs of the people. According to the fair share principle, given that the problems and suffering caused by a disease are tractable, global pharmaceutics research resources should be allotted so that they agree with the ratios of human suffering the diseases bring about. Now one of the diseases that have drastic consequences for millions is malaria. According to Flory's and Kitcher's calculations the amount of research directed to malaria was somebreak $ 85 million whereas according to the fair share principle it should be in the region of $1.75 billion (Flory and Kitcher 2004, p. 39).

Exactly how the figures are calculated is no concern here, but the very fact that the questions have been raised, and that normative and global issues of how research funds should be allocated indicate that very strong moral, legal, social, political principles that enter the scientific agenda. Needless to say, most of the pharmaceutics research is conducted within the pharmaceutical companies, tied of course with increasingly strong bonds to universities, publicly funded research centers and national institutes, US and European funding organizations etc. And almost needless to say, all of the stakeholders mentioned in the MASIS report, from Churches to NGO's to patient organizations want to have a say.

Here we can note yet another development in the recontextualisation of science, the emergence of synthetic biology. Synthetic biology is in interdisciplinary enterprise that combines biological, chemical, physical and engineering expertise to form standardized biological parts or components. These components – analogues of electric circuits and the like found on the shelves of electronics component shops such as Radio Shack – are then used to build or reconstruct biological devices, systems, or entire chromosomes. It is a philosophical interesting science since it does not aim at knowledge as such – not even applied knowledge – but rather biological devises, that is, things. Conceptually it is therefore closer to engineering than basic or applied science. Synthetic biology is expected to be of huge importance in e.g., the search for new energy sources – or medicine.

One of its success stories started in 2003 when Jay Keasling (from Berkeley) rewired a novel metabolic pathway for an artemisinin precursor, aretemisinic acid. He and his group combined ten genes from three types of existing organisms, namely plants, bacteria, and yeast into a platform called or bacterial chassis. These genes, orchestrated to work together, then produce enzymes that are able to turn acetyl coenzyme A into artemisinic acid. The reason why these advances are so

remarkable is not just the way scientists have advanced understanding but that fact that artemisinic acid is used in the production of artemisin, a malaria drug (see Lentzos 2009). Thus one of the hurdles on the way to a cheap malaria drug was removed. The neglected disease problem is a problem in part because pharmaceutical companies are only willing to invest in research and development for a particular disease if the potential patients have enough purchasing power to be able to by the drugs. Although malaria drugs have been researched this research and the production of artemisin has been colossally expensive. Malaria affects millions of people each year but their combined purchasing power has not been enough to be able to cover the costs. Now that Keasling's group has managed to scale up the laboratory results to a feasible industrial process the average cost of aremisin therapy has been reduced to less than 10 $. The stakeholders in this success story have involved not just Keasling's group and research institutes but also non-profit drug companies such as OneWorld Health as well as private foundations such as Bill and Melinda Gates Foundation.

8 Concluding Words

It is now rather universally accepted that inquiry in general and scientific inquiry in particular is a social affair, whether understood synchronically or diachronically, in that in scientists must rely on the testimony of others. It is also universally accepted that science is socially construed in the sense that concept formation, theory formation, experimentation and all facets of science are a result of cooperation and even negotiation. Whereas Ian Hacking raised the question "Social Construction – of What?" I have attempted to shift focus to "Social Construction – by Whom?" The result is not a historical review of the science/society interface but a modest attempt to highlight who the actors on the scene have been. One of the answers referred to the Gentlemanly culture of science of Seventeenth century England, another one to the emergence of the modern ideology – indeed "imaginary" of science as the business of autonomical communities of inquirers in the pursuit of truth. This imaginary was forged into an explicit social contract for Big Science (and Small Science followed suit) during WWI. As more recent developments indicate this social contract, indeed treaty, has now been dissolved. We have entered a period where a gamut of potential actors want to have their say in the governance of science.

References

Bacon, F. 1878. *Novum organum*. Oxford: Clarendon Press.

Berger, P., and T. Luckmann. 1967. *The social construction of reality: A treatise in the sociology of knowledge*. Garden City: Doubleday.

Burr, V. 1995. *An introduction to social constructionism*. London: Routledge.
Collins, H. 1983. An empirical relativist programme in the sociology of scientific knowledge. In *Science observed: Perspectives on the social study of science*, ed. K. Knorr-Cetina and M. Mulkay, 85–114. London/Beverly Hills/New Delhi: Sage.
Fleck, L., T.J. Trenn, and R.K. Merton. 1979. *The genesis and development of a scientific fact*. Chicago: University of Chicago Press.
Flory, J., and P. Kitcher. 2004. Global health and the scientific research agenda. *Philosophy and Public Affairs* 32(1): 36–65.
Foucault, M. 1989. *Order of things*. London: Routledge.
Hacking, I. 1999. *The social construction of what?* Cambridge, MA: Harvard University Press.
Irvine, J., and R. Martin. 1984. *Foresight in science: Picking the winners*. London: F. Pinter.
Kitcher, P. 2001. *Science, truth and democracy*. Oxford: Oxford University Press.
Knorr-Cetina, K. 1983. The ethnographic study of scientific work: Towards a constructivist interpretation of science. In *Science observed: Perspectives on the social study of science*, ed. K. Knorr-Cetina and M. Mulkay. London/Beverly Hills/New Delhi: Sage.
Kuhn, T.S. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Lentzos, P. 2009. Synthetic biology in the social context: The UK debate to date. *BioSocieties* 4: 303–315.
Merton, R.K. 1982. *Social research and the practicing professions*. Cambridge, MA: Abt Books.
OECD. 1966. *Ministerial meeting on science: Fundamental research and the policies of governments*, 2nd ed. Paris: Organisation for Economic Cooperation and Development.
Peirce, Ch. S. [CP] 1931–1935 and 1958. *The collected papers of Charles Sanders Peirce*, vols. 1–8, ed. C. Hartshorne, P. Weiss, and A. Burks. Cambridge, MA: Belknap Press. Referred to as CP.
Pettit, P. 1996. *The common mind*. Oxford: Oxford University Press.
Polanyi, M. 1962. The republic of science: Its political and economic theory. *Minerva* 1: 54–74. Also available at http://www.missouriwestern.edu/orgs/polanyi/mp-repsc.htm. Accessed 28 May 2013.
Reiss, J., and P. Kitcher. 2008. Neglected diseases and well-ordered science. Technical report 06/08, Centre for the Philosophy of Natural and Social.
Searle, J.R. 1995. *The construction of social reality*. New York: The Free Press.
Shapin, S. 1994. *A social history of truth: Civility and science in seventeenth-century England*. Chicago: The University of Chicago Press.
Shapin, S., and S. Schaffer. 1985. *Leviathan and the air pump: Hobbes, Boyle and the experimental life, including a translation of Thomas Hobbes, dialogus physicus De Natura Aeris*. Princeton: Princeton University Press.
Sintonen, M. 1990. Basic and applied research: Can the distinction still be drawn? *Science Studies* 2: 23–31.
Stokes, D.E. 1994. Science: The endless frontier as a treatise. In *Science, the endless frontier 1945–1995. Learning from the past, designing for the future*. Columbia University, Part I, 9 December.
The MASIS Report. 2009. *Report of the MASIS Group: Challenging futures of science in society – Emerging trends and cutting-edge issues* (Monitoring Activities of Science in Society in Europe), The European Commission.

Is Social Constructivism Soluble in Critical Naturalism?

Daniel Andler

1 Introduction

It has been recognized for a long time that naturalism is a blanket term covering a large array of philosophical positions and problem situations. Social constructivism, on the other hand, while having a seemingly much shorter history and encompassing a far narrower set of positions, also comes in a rather large variety of versions. To further complicate matters, of many stances or schools in philosophy, it may be said that they are, in some respect, naturalistic, or constructivist, or both, without claiming to be fully developed forms of naturalism or social constructivism, as the case may be. For all these reasons, it doesn't seem worth asking whether naturalism and social constructivism are, as such, incompatible, as appears obvious to many contemporary observers, or compatible, as others claim, or something in between. However another, more specific and, it is hoped, more instructive question may be asked today. On the one hand, the more general debates concerning social constructivism, either internal to the constructivist camp, or pitting that camp against the mainstream in analytic philosophy, have led to a somewhat streamlined and more focused position. On the other hand, naturalism has undergone a dual process of restriction and expansion: the first process, which has unfolded in the last half-century, has resulted in a fairly well circumscribed thesis, scientific naturalism, which is almost universally seen as opposed to social constructivism; the second process has led to a family of positions that try and amend scientific naturalism while keeping true to its basic inspiration, remaining in the same ballpark, and thus deserving, or so they hope, the label "liberal naturalism". The question worth asking

D. Andler (✉)
UFR de Philosophie, University Paris-Sorbonne,
1, rue Victor Cousin, 75005 Paris, France
e-mail: daniel.andler@ens.fr

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_19,

then, especially in the light of some of the motivations often put forth by defenders of liberal naturalism, is whether the streamlined, contemporary version of social constructivism can be accommodated within liberal naturalism.

Now liberal naturalism has in no way attained homogeneity or stability: it remains at present more of an editorial concept than a clearly established position, and I should not presume to speak for the eminent and sometimes diverging philosophers who have at least provisionally accepted the label. On the other hand, I agree with what many among them have to say, and have developed my own version of liberal naturalism, under the label "critical naturalism" – the label isn't new and what it refers to may not be that new either. Still, it is the position, in fact the only position that I find tenable in the naturalism debate, and I will take it as my reference point in my discussion of social constructivism.

2 Social Constructivism

What do I mean by social constructivism? Before we get into more conceptual issues, we need to clear the lexical ground. First, in some circles, a sharp distinction is made between constructionism and constructivism; also, some authors will use only the first term, others only the second. As I will use it, constructivism is indistinguishable from what some authors, such as Ian Hacking (1999), call constructionism: it refers to a doctrine whose centerpiece is the idea, whatever that idea precisely consists in, that something essential in the products or in the ontology of the sciences is socially constructed.

Second, I do not use social constructivism as a metonymy for the entire field of social studies of science, or of knowledge, nor as a synonym of relativism, or irrealism, or rejection of science. It is, to repeat, the theory, or any theory, that takes as its starting point the above-mentioned idea: such things as cells, quarks, dementia, laser beams, chemical compounds, and/or the theories wherein they figure, are in some sense socially constructed, rather than being just "out there" for us to discover or establish.

Note that this construal tends to rule out of bounds, or sweep to the outermost region of social constructivism, what some authors regard as the field's founding text (Berger and Luckmann 1966) *The Social Construction of Reality.* If, as one would gather from the title of that well-known book, all of reality, or everything in reality, is actually socially constructed, then, or so it would seem, nothing special is implied about the entities that are scrutinized by science. Still, and however that be, some of Berger and Luckmann's inspiration has persisted as an undercurrent in later social constructivism, as Jeff Coulter reminds us (Coulter 2001): the Marxian conceptual apparatus was deployed as an antidote to what the authors saw as an illegitimate, self-interested and all too successful attempt of the ruling classes to pass off as objective facts regarding an out-there, immovable nature, what is in fact the product of social processes of exploitation and alienation. In particular, entities which appear, in such a set-up, as real objects, comparable to rivers or stones, are unmasked

as being mere constructions: the "critique of reification (hypostatization, or the fallacy of misplaced concreteness, to name two other common variants of the same concept)" was, and remains, a guiding principle of much constructivist thought. So that "the demonstration that some phenomenon was 'socially constructed' was a major methodological procedure in the process of its dereification" (*id*., p. 83).

This critique of science is in fact of the essence of constructivism; what is subject to variation is the degree to which science is allowed to survive the critique. To a first approximation, the radicals mean to discredit science altogether, the moderates merely want to reduce it to size, and the two poles subtend a continuum. But both radicals and moderates subscribe to three permanent strands, which are already clearly apparent in the founding stage: the unmasking of spurious inevitability, the social-ethical concern for improving the lives of the oppressed, the crucial yet hidden role of language in the operation leading to the present situation.

Following Ian Hacking, we can unpack these further. First, we can peel apart the pragmatic (or praxeological) and the theoretical dimensions. Hacking urges us to look beyond the theoretical claim that X is socially constructed, and ask what purpose underwrites such a claim; the answer is some variant of the following conjunction : "X is bad as it is now", and (indeed) "We would be much better off if X were done away with or at least radically transformed" (*id*., p. 6). This pragmatic goal is then subserved by a composite, properly theoretical claim:

> X is epistemically not inevitable, while it is presented by science, or by supporters of science, as inevitable.

By "epistemically inevitable" I mean, still following Hacking, that it *could not* have turned out that X would *not* figure in our epistemic situation–or, cancelling out the double negation, any possible epistemic situation would necessarily include X. An essential part of the claim (hence the qualification "epistemic") is that the modals "could" and "necessarily" refer not to the contingencies of scientific discovery, funding priorities, administrative and political decisions, vagaries of scientific geniuses, etc., but to the idealized development of science. The social constructivist holds that X is a contingent aspect of science as we know it today, and that science could have turned out, even in ideal circumstances, and may turn out tomorrow, to be such as not to include X (*id*., pp. 69–80).

Now what sort of thing is X supposed to be? It can be one of two things: a fact, theory, equation, explanation, more generally, an *idea*, or an entity. In the first case, the inevitability thesis takes on the more precise form given to it by Hacking under the label *contingency*. His prime example is Pickering's thesis that quarks are socially constructed, and he summarizes it thus: "There could have been a research program as successful ('progressive') as that of high-energy physics in the 1970s, but with different theories, phenomenology, schematic descriptions of apparatus, and apparatus, and a different, and progressive, series of robust fits between these ingredients. Moreover [. . .] the 'different' physics would not have been equivalent to present physics" (*id.*, p. 72).

Note that X, in the preceding example, is the *idea* of a quark, or the *thesis* that quarks, with their attendant properties, are a constituent of reality. But X can also be

an *entity*, more particularly a class or kind. Here, according to Hacking, the social constructivist espouses a form of *nominalism*. She rejects anything resembling the notion that science, however high-grade or ideally completed, could be, up to idealization, isomorphic to nature; Hacking coins a word for this idea, which he wants to insulate from the wider debate on scientific realism: he calls "structure-inherentism" the thought, rejected by constructivism, that nature has a structure of its own, which science aims with some success at recovering or representing.

At this point, social constructivism has thoroughly undermined the *authority* of science: if science *could have* been otherwise, stating other facts about other things ordered in different arrangements, then in its present, or for that matter future state, it cannot be regarded as having *the last word* on anything, even though it can still be seen as providing a respectable guess at what can be said regarding certain topics of theoretical or practical interest. The most interesting contributors to social constructivism are not tempted by a blanket debunking of science, in fact they often profess a genuine admiration for science. What they do not espouse is what they see as the exaggerated deference towards science deemed rationally, as well as politically correct in polite society. And the basic reason why this deference is not warranted, according to them, is that it is based on the false premise that nature "dictates" the big book of science. That book, they insist, is the product of social constructions, doubtless operating in conjunction with the resistance afforded by the world "out there", but not guided towards a unique endpoint. Or to put it into an extreme form, the world does *perturb* the social processes of science, it does not *determine* them even in principle – which is exactly the reverse of the classical rationalist picture: the social processes *perturb* the interactions between world and science, they do not *determine* them even in principle.

3 Naturalism: Classical and Critical

It is often said these days that naturalism is, as I have put it in previous writings, by reference to Sartre's characterization of Marxism, the "unsurpassable philosophy of our time". Almost every philosopher, according to this view, subscribes professionally to naturalism (although she might also be a devout church goer, for example). However popular, this view is based in part, I believe, on a conflation of two senses of "naturalism".

In the loose, non-technical sense, naturalism refers to an attitude, a decision or a commandment: the naturalist intends to look at things as they are, that is, as they appear in their actual, everyday existence, with no details deemed too minor or trivial as to be neglected, and with no regard for preconceived ideas of what they are, let alone norms prescribing what they should be. The technical or philosophical sense of naturalism (often referred to as scientific naturalism), on the other hand, has to do with a certain conception of nature: as circumscribed by the ontology of the

natural sciences, as epistemically accessible exclusively through the methods of the natural sciences, and finally as comprising everything that exists. Sitting somewhere in between the loose and the technical sense, there is a sense of naturalism as a conception of the nature of philosophy: philosophy is said to be "continuous" with (natural) science. This metaphilosophical naturalism partakes of both basic construals, insofar as the sciences are, by vocation, the foremost scrutinizers of things as they really are, so that anything continuous with the sciences will be naturalistic in the loose sense; while on the other hand, subscribing to such a view of philosophy is consonant with a conception of the natural sciences as having privileged access to nature hence to the whole of reality.

Why naturalism seems to some unsurpassable may be due to the historical fact that very few philosophers today still believe that they can rely exclusively on conceptual analysis and a priori theorizing: most of them appreciate the need to carefully consider things as they are, whether presented through commonsense, everyday language and practices, or, of course, through the sciences and non-theoretical bodies of expertise. In other words, most contemporary philosophers partake of naturalism in the loose, non-technical sense. As for the technical sense, there remains some leeway: in fact, we are witnessing what may be a change of mood. Nonetheless, at the present juncture a large number, definitely the majority of English-writing philosophers, are drawn to (one or another version of) naturalism in the technical sense. As for the hybrid sense (the continuity thesis), it is flexible enough to gain acceptance by many philosophers: being "continuous with science" may be taken to be no more than, attention to detail, careful argumentation, parceling out of large problematics into manageable chunks, distributed within the community of professionals, and naturally the full cognizance of the results of science – nothing which doesn't agree with the methodology of analytic philosophers, from Aristotle onward.

For present purposes, I will take naturalism to mean scientific naturalism augmented by the continuity thesis, ignoring further specifications (such as a stand on physicalism). As mentioned above, it is generally agreed on both sides that social constructivism is incompatible with naturalism as understood in contemporary philosophy: the main arguments will be summarized shortly. The possibility that remains is to investigate whether by liberalizing naturalism in one way or another, one might not make room for social constructivism or at least a sizable chunk of it. And again, there are disparate proposals for the liberalization of naturalism, so that there may well be no definite answer to the question: in fact, that very question may turn out to serve as a benchmark for grading varieties of liberal naturalism. Still, I think it fair to say that all forms of liberal naturalism share the following features. On the positive side, just like strict naturalists (i) they reject supernaturalism; (ii) they have a high regard for science in general, and natural science in particular. On the negative side, unlike strict naturalists, (iii) they harbor doubts about the cogency and/or eventual success of (some or all) the currently fashionable "naturalization"

programs in philosophy of mind, action or culture, in ethics or esthetics[1]; (iv) their respect for natural science does not go so far as to grant it "universal coverage" – that is, they do not believe that natural science has the potential to deliver all the relevant general facts regarding the entirety of reality.

My own brand of liberal naturalism, which I call "critical naturalism", arises from an attempt to find an acceptable answer to the following pragmatic concern: what is the proper *attitude* towards ongoing naturalistic (or naturalization) programs, whether scientific or philosophical (insofar as there is a relevant distinction in such borderline matters)? The "classical" naturalist takes it as rationally, as well often as ethically, obligatory to support such programs unreservedly, since they represent current science's best attempt at an enterprise that is *bound* to succeed (sooner or later, and not necessarily as a direct result of the ongoing programs: science is known to backtrack often). The anti-naturalist sees such programs as thoroughly misguided, a waste of time and resources and possibly a betrayal of the scientific spirit, one of whose aspects is a keen sense of the limits of science. The non-naturalist is simply skeptical or ironic, withholding his prognosis of the outcome of these various programs. My critical naturalist, on the other hand, lends them critical support: he believes they should be taken seriously, he harbors the hope that important things will be learnt on the way, but sees no reason to believe they are *bound* to succeed, and regards it as his duty to submit them to close scrutiny, as past experience has taught him that premature or misguided attempts, pursued in the name of naturalization, would have been curtailed or redirected for the better if philosophers and scientists had been less gullible. Nor is his caution necessarily based exclusively on a negative judgment on the record of previous attempts, or on the state of some of the ongoing programs (although such a negative judgment may play a role). He is inclined to think that science has limited scope, not only inasmuch as it doesn't, by a long shot, "cover" everything, but also because even when it does colonize a fragment of reality, it seldom if ever provides a complete account of that fragment: it remains forever incomplete, with bits and pieces of knowledge applying to bits and pieces of the world. Moreover, when it comes to human affairs, he tends to believe that norms of all kinds, bound to local circumstances, thoroughly permeate the thoughts and behavior of agents in such a way that the natural constraints which are at work are rarely sufficient to construct reliable accounts, let alone predictions. But again, although he comes close to certain themes dear to anti-naturalists, he remains true to the naturalistic spirit by granting a central importance to the study of these natural constraints.

[1]This emphatically does not mean that they necessarily oppose all of these programs, or consider them as clearly hopeless or already bankrupt: on the other hand they are far from thinking of them as *bound* for success. However they do not think of the whole lot of them as necessarily misguided, as this would presumably make them into thorough-going anti- or at least non-naturalists.

4 Simple Attempts to Reconcile Social Constructivism and Naturalism

I have been discussing the various senses of naturalism up to now as if they were relevant only to philosophers. But we must now consider their relevance for the social constructivists themselves. Some of them – mostly among those who we philosophers tend to read – are philosophers, by training at least and sometimes by current affiliation. But almost all of them think of themselves either in part or exclusively as social scientists sociologists, historians, anthropologists, geographers . . . I am ruling out of bounds the entire crowd of non-scientists, by which I mean scholars in the humanities who do not regard themselves as scientists in any but the loosest sense – they do academic research – and who, more often than not, harbor a strong aversion towards science. For those social scientists of constructivist bend, what does naturalism stand for?

This is where failure to distinguish different senses or kinds of naturalism is bound to lead to confusion. For in the loose, non technical sense, social constructivists are, if I may say, quintessential naturalists: naturalism is their *raison d'être* and their trademark. The post-Kuhnian collapse of what Philip Kitcher has called the "Legend" (Kitcher 1993), the idealized picture of science as a purely rational enterprise stabilizing gradually toward a well-founded, logically structured and unified theory, opened the way for the study of science as it *really* is, not science as philosophers of science and some of their famous scientific friends would like it to be. For some philosophers and historians of science, this was the beginning of so-called descriptive philosophy of science, with the attendant project of "integrated" history and philosophy of science (&HPS)[2] and the de-emphasizing of general philosophy of science in favor of philosophy of particular sciences. For those, philosophers and sociologists, who had been raised in a Mertonian view of the place of sociology, the new frontier was the study of the *contents* of science (of the products as well as the processes of science) as social phenomena, as effects of social forces. Rather than reconstructing an ethereal counterpart of science in the non-natural realm of ideas (in Popper's World 3, for example [Popper 1972]), the new sociologists wanted to know how science in the natural world is naturally produced, and the founders of SSK had good reasons to think that the proper level at which to attack the natural production of science is social. All of this is well-known. Less so is the sequel: today philosophers of science with a naturalistic bend (in the loose sense) are happy to co-operate with some of the descendants of SSK, practitioners of various strands of what Derek J. de Solla Price (1961, p. 128) proposed to call "the scientific humanities", within the broad framework of a new

[2] "The founding insight of the modern discipline of HPS is that history and philosophy have a special affinity and one can effectively advance both simultaneously". From the website of &HPS: http://www3.nd.edu/~andhps/about.html

interdisciplinary field called "social epistemology", and the attendant projects of feminist philosophy of science and of "socially relevant" philosophy of science.

Are we to conclude that social constructivism and naturalism are now comfortable bedfellows? Seen from the immense distance provided by the history of philosophy, it may be the case – as Finn Collin has argued: constructivism, he claims, is but the latest manifestation of "the trend toward naturalism": "We need to view constructivism and the new science studies in which it is embedded as the latest phase of the movement within Western thought toward an ever more pervasive naturalism. By naturalism I understand the contention that everything in the universe, including man, is just a part of material, empirical nature" (Collin 2001, p. 425).[3] Constructivism aims at a "theory of science [which should] be conducted in a genuinely empirical spirit" (*id.*, p. 430), and thus partakes of the naturalistic turn according to him.

But this does not do justice to the dialectical situation at all: now is the place to remind ourselves of the bones of contention. For naturalism in the technical sense, which is the preferred reading of this statement (again, this is the difference between the adjective and the noun: a naturalistic–equivalently: empirically-minded – philosopher of science need not be a defender of naturalism in the technical sense), cannot countenance an account that keeps the social realm separate, as a province of ontology and as an agency. To the naturalist (*stricto sensu*), the social is part of nature, social processes are natural processes, with causal powers reducible to natural causation, nature is what the natural sciences talk about. This is not acceptable to most social constructivists, who will argue in favor of the reverse dependence: nature as the domain of the natural sciences is a product of social forces. Moreover, the classical naturalist is not about to let go of a basically realist picture in which nature impugns on the scientific process, in such a way as to nudge it towards (if not directly impress upon it) a true account. This, to the social constructivist, is the ultimate naïveté.[4] Moreover, the contingency of social forces implies the contingency of the outcome: science could be other than it is at present, a relativist belief which is unacceptable to the true naturalist. Finally, social constructivists tend to accept the traditional "bifurcationist" view according to which the social sciences deal with cultural entities and processes that are out of reach of the natural sciences.

Are we then back on square One, just reiterating that naturalism and social constructivism disagree? Not quite: we are now clear on the non-technical sense of naturalism that allows for peace-making (the process is in fact enduring), and setting this issue aside, we can attack the deeper one: Can we soften classical (scientific) naturalism so as to accommodate some recognizable form of social constructivism?

[3]Collin has recently published a book (which I have not read) in which he develops the point: *Science Studies as Naturalized Philosophy*, Dordrecht: Springer, 2011.

[4]Not only of course to the social constructivist: Hacking for example approves Latour and Woolgar (1979) for asserting, in essence, that "We should never explain scientific beliefs by facts or truths, except as a kind of shorthand" (Hacking 1999, ch. 3).

As a preparatory step I will now briefly review two recent attempts at showing, *in media res* rather than through abstract arguments, that accommodation is possible.

They are both is due to some true, strict naturalists, philosophers and biologists working in evolutionary psychology. Their field is engaged in a never-ending battle with the anti-naturalist majority in anthropology, sociology and related social sciences. Within that camp, the most vocal group is made up of social constructivists, understood for the purposes of this particular debate in a wide sense (social constructivism in the strict sense being something like their theoretical core): they include all those who are "commit[ted] to the idea that individuals and societies have enormous flexibility in what they can become, in contrast to the inflexibility and determinism attributed to evolutionary approaches to human behavior" (Wilson 2005, p. 20). Between them and the evolutionary psychologists, whom I am casting here as the vanguard of current scientific naturalism, "[the debates usually become so polarized that they reveal the worst aspects of tribalism in our species. Each side regards the other as the enemy whose position has no substance or rational basis, other than being ideologically driven" (*ibid.*). D.S. Wilson recently, and a few years back philosophers R. Mallon and S. Stich (2000), have proposed a truce, based on the idea that rather than two irreconcilable worldviews, at their best the two sides pursue, unwittingly, complementary research programs.

Mallon and Stich take the emotions as their testing ground. Naturalists, following Darwin, believe they are in some sense innate and universal. Constructivists stress the immense variability, across epochs and cultures, of their behavioral expression, of their semantics, of their taxonomy; they go as far as saying that the repertory of emotions varies from culture to culture. Both sides can boast strong evidence, and keep, unsurprisingly, unearthing more. However, argue M&S, they are misled by the word "emotion", or rather, they subscribe to a philosophical theory of meaning and reference that creates the appearance of a conflict, and can easily be discarded. For brevity's sake, the argument may be simplified thus:

1. Both sides actually converge on a hybrid model of the emotions, involving both cultural, context-bound components and innate dispositions. Roughly speaking, the innate dispositions tend to favor certain transition patterns, from perceived situation ("antecedent conditions") A to emotion ("recruited response tendencies", which include a proto-emotion activating in turn a specific subjective experience together with a number of "programs" regulating facial muscles, etc.) to behavioral ("measurable") response C. In this tripartite model, due to Levenson (1994; I quote from Mallon and Stich's paper 2000, pp. 143–144), the middle segment is naturalistic, the first and last segment, respectively "appraisal" and "display and feeling rules" are under the control of cultural learning. The details are contentious, for sure, and should the two parties decide to actually co-operate, quite a number of their respective findings would have to be seriously re-interpreted. No insurmountable obstacle however would seem to get in the way of a reconciliation.
2. The bone of contention is universality: one side thinks of the emotions as universal, the other believes there is overwhelming evidence that they are not.

> Whatever the merits of the above model, it will clearly not lead either camp very far on the path to peace as long as they disagree so thoroughly on this point. Mallon and Stich's suggestion is to distinguish "thick" and "thin" concepts of emotions. Thick concepts involve links to large bodies of folk psychology, which anchor a given emotion, say anger or shame, under those names, in the American culture, to a host of dispositions, situations, traditions . . . : a given instance of an emotion is, say, anger, or shame, only if it is related in the prescribed way to each one of the corresponding set. As a result, it is indeed near impossible that an emotion in another culture would exactly fill the same role, granting the implausible assumption that some analogs could be found for each of the dispositions, situations, traditions . . . to which the original emotion is related in the American culture. *Ergo,* on a thick conception of how to define the emotions, they are clearly *not* universal. On a thin conception, one which severely limits the set of folk-psychological entities allowed to functionally constrain the individuation of the emotions, on the other hand, it is much easier for a Japanese, Sicilian or Inuit emotion to fit the bill of American anger or shame, as the case may be: on that reading, emotions (some emotions at least) could well be universal.

M&S's conclusion is that the fundamental disagreement regarding universality can simply be dissolved by letting go of a certain way of conceiving of the individuation of emotions. Alternatively, one camp may prefer to introduce a shadow vocabulary for the emotions: while, say, constructivists would hold on to the usual, culture- and language-dependent repertory of emotions, from anger and shame to the Japanese *amae* or the Ifaluk *song*,[5] the evolutionary psychologists would focus on a sparser repertory of "core emotions" – "core anger", "core shame" etc., that would, in different cultural contexts, give rise to local repertories of (the other sort of) emotions.

This proposal follows a very general strategy first introduced by Chomsky: the language capacity, the ability to acquire the full mastery of a language, is universal, although languages are many and diverse. In other, broader terms: nature sets the rules, culture and history "interpret" them, or (as Chomskyans would put it, fix the parameters). I will not attempt to evaluate Mallon and Stich's attempt at conciliation on its own. To me, the main question is how far this strategy, even if successful, will take us. Mallon and Stich express the hope that the case of emotions is paradigmatic. I would not bet on it: although they are correct in stressing that constructivists as well as naturalists see the matter as "one of [their]s success stories", with hindsight it looks like a rather easy case for the naturalists and a hard case for the constructivists

[5]Following Lutz (1988), Mallon and Stich (*id*., p. 139) characterize *song,* a concept belonging to the Ifaluk, a Micronesian people, as an emotion "akin to anger", yet comprises "a strong moral component" that is not part of our concept of anger. *Amae,* write Mallon and Stich (*id*., p. 145), following Morsbach and Tyler (1986), "is a Japanese emotion that is unknown (or at least unrecognized) in the West". According to a Japanese psychiatrist whom they quote from secondary sources, *amae* is characterized by "a sense of helplessness and the desire to be loved" (*ibid*.).

(hence their insistence on winning that one: if they can prevail there, the field is theirs, or so they feel). On the face of it, constructivists are on firmer grounds the further one moves away from basic functions: science, which is our concern, or for that matter theoretical knowledge in general, are anything but basic, they don't seem universal in anything like the way language, emotions, survival and reproduction behaviors are, and they do seem to involve thick and interspersed layers of "culture" and "nature".

Wilson's proposal (2005, pp. 20–37), the second that I propose to examine, is more ambitious than Mallon and Stich's, and tries to reach deeper into the constructivist's soul. For lack of space I will merely list the key ingredients of his argument. First, Wilson reminds us that not all evolutionary psychologists subscribe to the narrow, better publicized, version of their basic evolutionary tenet: according to that popular version, the mind is a collection of specialized adaptations dating back to our ancestral environments (Tooby and Cosmides 1992). Some evolutionary psychologists believe that "there is more to evolution than genetic evolution", and Wilson puts considerable stress on a parallel with the immune system, a well-studied "example of a physiological evolutionary process". Some also further emphasize that "there is more to evolution than adaptation": important features of contemporary human culture may have piggy-backed on traits which were selected for, and to which they are now but distantly connected. Second, as I indicated earlier, Wilson identifies at the core of the constructivist position the belief in the ample leeway that biological constraints leave for social and individual choice. He distinguishes a moderate and a radical version of this "flexibility thesis". The moderate version suspends judgment regarding the degree of freedom left to culture and history by the biological constraints; the radical version claims that this freedom is unlimited: "anything goes", any conceivable pattern is actually possible. Wilson claims that simultaneously enriching the basic evolutionary picture by admission of non-genetic evolution, and of non-adaptive evolution, and restricting the flexibility thesis to its moderate version, makes it possible to bring them together. The key is to remove the impression that evolutionary theory cannot accommodate the "open-endedness" which constructivists, and Wilson himself, see as characterizing the human world. This impression has solidified in many a mind in the form of the notorious blank slate metaphor: we are made to believe that only one box can be ticked among the following two:

Box 1: The mind is a blank slate on which culture, history and personal experience are free to write anything.

Box 2: The mind has a structure that essentially predetermines the main features of the content that culture, history and personal experience impress on the mind.

Wilson rejects the dichotomy. The blank slate is a metaphor with real, albeit limited validity: "[It] might be a total failure as a mechanistic conception of the mind but still be perfectly valid with respect to the open-ended nature of individual and societal change. [...] There is a difference between *potential* for individual and societal change and *equi-potential*. If by blank slate we mean "anything can be

written with equal ease", then that part of the metaphor is false. [...] fulfilling the valid aspects of the blank slate metaphor requires abandoning the invalid aspects. Potential does not mean equi-potential" (*id*., p. 28). Wilson goes on to illustrate the integrative power of evolutionary *cum* cultural thinking by sketching a theory of stories as being, in some sense, "gene like", and serving the function of adapting humans to their current environments; narratives are more central to the constructivist worldview than, say, emotions, and by granting them an important role, and showing that the most naturalistic among naturalistic programs, *viz.* evolutionary psychology, can countenance them, he proposes a deeper alliance than do Mallon and Stich.

I have given Wilson's attempt short shrift, but I hope to have conveyed three important ideas: (i) There exist serious, constructive attempts to combine in a principled way some central constructivist insights and a naturalistic inquiry. (ii) The first step, in all such reconciliations, consists in devising a model that does justice to the role of the natural endowment (the "fixed part" of the mind) and to the role of the social, historical, biographical context (the "variable part" of the mind). (iii) By itself, this first step goes no further than prepare the ground for a true meeting: it provides an enabling condition by removing a seemingly insurmountable contradiction. In order to address the real concerns of the constructivists, one must take on board at least some of their initial motivations (as Wilson has done by granting full acknowledgement to the thought of people and cultures having a free hand in shaping their destiny) and be willing to enter the heart of their ontology (as Wilson has done by focusing on stories).

5 Refocusing on the Original Problem

Still, the feeling remains that even Wilson's attempt leaves the issue unresolved. The reason is that the problem as he states it appears to be somewhat different from our original problem, although they are certainly related. Our problem (Problem A) concerned social constructivism as a doctrine within science studies bearing on the status of scientific theories and of the entities they refer to. Both Wilson and M&S are concerned with (Problem B) the constituents of human minds and cultures: are they "natural", i.e., is there reason to think that they can be accounted for by the natural sciences with their present conceptual apparatus or a reasonable extension thereof, or are they "constructed", i.e. shaped by the history of local interactions in human groups, a mix of collective intentionality and the blind influences of physical and man-made environments?[6] In a sense, Problem B is an instance of Problem A:

[6]The reader is perhaps reminded of the distinction made by Michael Bradie between what he calls the Evolution of Epistemological Mechanisms (EEM) and the Evolutionary Epistemology of Theories (EET). See e.g. Bradie (1989) or the *Stanford Encyclopedia of Philosophy* entry "Evolutionary Epistemology".

the natural sciences of the mind and culture, seen from a naturalistic vantage point, claim to have identified certain classes of entities and processes in the world, while social constructivists insist that the concepts these sciences use, and (in a sense) the entities they refer to, are the product of social forces – in particular, they could have been other than what they are (contingency), they do not mirror the structure of the world-in-itself (nominalism or rejection of "structure-inherentism") and the sciences themselves don't have the last word on these matters (anti-authority).

Yet in another sense, Problem A is an instance of Problem B: science is, after all, a constituent, manifestation, or product of human minds and cultures. If Problem B is solved along naturalistic lines, that solution will carry over to Problem A, *but not in the expected way*, that is, not as both traditional philosophy of science and social constructivism in the Strong Programme orientation view it, by showing how social forces shape the content of scientific theories. The natural sciences (and presumably the conceptual bricks out of which they are made), and hence the sciences in general, would instead be shown to be the outcome of natural processes. This is, of course, Quine's idea of naturalized epistemology, with its attendant circularity: science would show (by the usual rational means) that it is in fact the outcome of entirely natural (rather than rational) processes (or to put it in Quine's own terms, epistemology would turn out to be "a province of psychology"). If, on the other hand, the solution to Problem B is constructivist, it immediately carries over to Problem A. We would be assured that science, a product of mind and culture, is not an object for natural science, but ... a product of mind and culture! This is not a tautology, for the first occurrence of "product" refers to the outcome of the process, while the second occurrence refers to its nature or constitution. There is a circle there too, though, which is the well-known principle of reflexivity familiar to the Strong Programme: social constructivism as a science of science affirms authoritatively, as against rival theories (classical normative philosophy of science and scientific naturalism), a picture of science which includes the fact that social constructivism itself does *not* have the authority to claim what it does.

This *is* confusing, but out of this confusion clarity emerges in just two steps.

First, *pace* both Quine *and* the Strong Programme, one cannot dispense with analytically distinguishing the epistemological framework in which the discussion is conducted and the object of the discussion. This move makes it possible to formulate the following theses:

(I) *Internalism.* Knowledge, and in particular science, are governed by constitutive norms that are in principle accessible to, though in actual fact more or less accessed by, the reflective mind operating in a critical, dialogical setting.
(EN) *Naturalistic externalism.* Knowledge, and in particular science, are governed by blind natural processes that trump the internal norms.
(EC) *Social constructivist externalism.* Knowledge, and in particular science, are governed by blind social processes that trump the internal norms.
(SS) *Scientific Success.* Natural science succeeds in its own terms: by following its own rules, it manages to provide adequate, objective representations of the world.

We can then consider the following six-cell logical space:

I ?γ → SS? ↓	Yes: I	No: EN	No: EC
SS : *yes*	Trad	Nat	Sr
SS : *no*	Sk	BG	SC

Out of six possible positions, only three are relevant for present purposes:

- *Tradition* (Trad). Natural science is successful, yet the norms governing knowledge and science in particular are autonomous and are the *raison d'être* of normative epistemology.
- *Naturalism* (Nat). Natural science is successful and has full coverage, including the scientific process itself, which is accessible to its empirical methods.
- *Social constructivism* (SC). Natural science is not successful in its own terms, and the scientific process is accessible to the empirical methods of social science.

Skepticism (Sk), the view that the quest for knowledge if subject to constitutive norms, yet fails, falls outside the scope of the present discussion: neither parties want to defend it. So do, by the same token, Serendipity (Sr), the view that although science is governed from the outside by social forces, nonetheless it succeeds in its own terms; and Blind Groping (BG), which sees our quest as governed by natural forces and failing.

The second clarifying step is to notice that there are three, not just two positions facing one another. Part of the confusion is due to the fact that alliances can shift: Nat and SC's natural adversary is Trad, as regards (I), while as regards (SS) Trad and Nat are both pitted against SC.

In this regard it is instructive to quickly examine yet another unsuccessful attempt to reconcile social constructivism and naturalism. In a recent paper, Melinda Fagan (2010) argues that naturalism should yield to social constructivism on the matter of justification: in the wake of the long debate concerning the possibility of maintaining the normative function of epistemology in a naturalized setting, Fagan's plea to the naturalist is to take stock of the fact that when the philosopher of science evaluates a research program, she necessarily does it from the standpoint of our present criteria, and those criteria are necessarily the outcome of our present historical and social circumstances. So by combining social constructivism with the generally naturalistic stance of contemporary philosophy of science, she proposes to save the project of a normatively potent naturalistic epistemology, in other words, to have the best of three worlds: the normativity of tradition, naturalism, and constructivism. However, I fail to see how one goes from the obvious fact that we reason with our present resources to the idea that we reason with socially constructed norms of justification, unless one is already convinced that norms and epistemic practices in general are socially constructed. I would think that much, much more evidence is required to reach that conclusion. This in fact is precisely the kind of evidence that social constructivists claim to unearth – their central idea, as Mallon (2007, p. 93) reminds us, is that "human decision and human culture exert profound and *often*

unnoticed influence": surely the fact that we reason with the means and tools at our disposal cannot go *unnoticed!* So Fagan's attempt at synthesis seems to fail.

6 Critical Naturalism and the Hard Core of Social Constructivism

Now we come at last to the title question: if classical or strict naturalism as we know it cannot absorb social constructivism, can *liberal*, and more particularly, *critical* naturalism do it? Does the move away from strict naturalism provide enough elbow room?

One feature of critical naturalism (CN) is its willingness to challenge natural science, not just locally like the normative philosophy of science that Fagan wishes to make secure (assessing competing research programs, present or past), but at a more global level. For one, universal coverage is as unwarranted an assumption, for a critical naturalist, as in-principle limitation as claimed by anti-naturalists. CN also encourages a certain form of irreverence towards natural science, taking the liberty of questioning, in some cases, the relevance or perspicuity of certain research programs with respect to the issues at hand, no matter their credentials *qua* science; this attitude is one that SC finds congenial. Last but not least, CN rejects, even as a regulative ideal, the concept of a completed, tension-free science. Together, these tenets of CN bear a certain resemblance to some of the themes and motivations of social constructivism. In particular, CN embraces scientific pluralism, a form of anti-absolutism which is an objectivist cousin of relativism as espoused by SC. From there, it is an easy step for CN to reject the idea that natural science forms a solid block facing another solid block, social science, to be either carefully preserved from naturalistic encroachment or to be on the contrary eventually absorbed by it. It then seems pointless to rule out hybrid accounts of the scientific process, in which both the natural sciences and social factors weighing on the process itself would enter.

At this point social constructivism of the less radical, pro-science variety, seems poised for absorption. For not only can it not avail itself of a metaphysical shield: the idea that the social and historical factors operating in the shaping of science are disconnected from the order of nature – that would run counter its naturalistic commitment. But as I have just shown, NC incorporates the critical spirit, and the attendant ideas regarding the empire of science, that were the initial inspiration of SC. Social constructivism would presumably not find this outcome initially at all likely or appealing. But the recent history of the movement, no longer united, and the dialectical turns of Bruno Latour, forever the moving target (see e.g. Latour 2003), suggest that the pressure is high. The only escape, I submit, is for social constructivism to bite the bullet and renounce any pretense at naturalism.

In his insightful review of André Kukla's (2000) and Ian Hacking's (1999) books (Rouse 2002), Joseph Rouse seems to suggest precisely the kind of move which would grant social constructivism immunity from absorption in naturalism,

however critical, and from collapse under the weight of its inconsistencies. As I read him, both Kukla and, in a considerably more sophisticated fashion, Hacking, fail to target the admittedly hidden core of true constructivism, or perhaps, the sort of revived social constructivism that Rouse, judging from the subtitle of his recent book (2003), would regard as counting as "naturalism reclaimed". Be that as it may, Rouse's insight is that almost everyone, including probably many constructivists a lot of the time, has tended to view the *content* of scientific theories as conceptually independent of the actual processes that select them. As he puts it, "science [according to the majority view, to be rejected] is conceived as an interface between human society and the natural world, and the question is whether what happens at that interface is best explained by the science's rational transduction of information from outside, or as only a reflection of 'factors' internal to human societies" (2002, p. 69). But the *meaning* of scientific theories, however the issue is settled, "can be discerned without drawing upon a background of practical skills, equipment, visual images, material surroundings, institutional networks, and discursive patterns" (*ibid*.).

Now suppose to the contrary that "language is only meaningful in the midst of extensive practical dealings with particular surroundings (often carefully arranged in laboratories, or the extension of laboratory practices into other settings), which thereby also acquire intelligible structure. Meaning would then be found neither in language by itself, nor in the world apart from language, but only within a-world-that-pervasively-incorporates-discursive-practices-and-norms" (*ibid*.). To put it differently, the suggestion is that for scientific language at least, there is no "cat on the mat" verification scheme: the meaning of a scientific "cat", a scientific "on" relation, and a scientific "mat" only emerges from worldly events involving the entities under scrutiny, the investigators, the equipment they use, the social norms governing their practice. There is no such thing as an investigation of the world aiming at determining whether or not the (previously determined) factoid "the cat is on the mat" is a genuine fact. To reiterate, meaning, on that view, emerges from the interplay of language and a world that is not just out there, but already permeated by human practices and norms.

If this is right, and social constructivism, in its "post-modernist" late phase, espouses that view, then it becomes clearly incompatible with any form of naturalism, however liberal; and concomitantly, it also sheds any remainder of its initial naturalistic tenet: it may still claim to shun the apriorism of traditional philosophy of science, but it ceases to be empirical in any straightforward sense. Social constructivism has found a firm anchor, philosophically less precarious than its Marxian defense of the exploited, or its suspicion of science, and rooted in its initial source of inspiration, viz. Wittgenstein's conception of language: the co-constitution of language, practices and the world is a strong philosophical view, well represented in current debates.

My purpose in this paper is not to defend that view, despite the grain of truth I think I can discern in this radical constructivism, nor to justify my adherence to critical naturalism, I have merely tried to show that constructivists can't have it both ways, nor can we.

References

Berger, P.L., and T. Luckmann. 1966. *The social construction of reality*. Garden City: Doubleday.

Bradie, M. 1989. Evolutionary epistemology as naturalized epistemology. In *Issues in evolutionary epistemology*, ed. K. Hahlweg and C.A. Hooker, 393–412. Albany: SUNY Press.

Collin, F. 2001. Bunge and Hacking on constructivism. *Philosophy of the Social Sciences* 31(3): 424–453.

Collin, F. 2011. *Science studies as naturalized philosophy*. Dordrecht: Springer.

Coulter, J. 2001. Ian Hacking on constructionism. *Science, Technology & Human Values* 26(1): 82–86.

de Solla Price, D.J. 1961, rev. 1975. *Science since Babylon*. New Haven: Yale University Press.

Fagan, M.B. 2010. Social construction revisited: Epistemology and scientific practice. *Philosophy of Science* 77(1): 92–116.

Hacking, I. 1999. *The social construction of what?* Cambridge, MA: Harvard University Press.

Kitcher, Ph. 1993. *The advancement of science: Science without legend, objectivity without illusions*. New York: Oxford University Press.

Kukla, A. 2000. *Social constructivism and the philosophy of science*. London: Routledge.

Latour, B. 2003. The promises of constructivism. In *Chasing technology: Matrix of materiality*, ed. D. Ihde and E. Selinger, 27–46. Bloomington/Indianapolis: Indiana University Press.

Latour, B., and S. Woolgar. 1979. *Laboratory life*. London: Sage.

Levenson, R. 1994. Human emotion: A functional view. In *The nature of emotion: Fundamental questions*, ed. P. Ekman and R.J. Davidson. New York: Oxford University Press.

Lutz, C. 1988. *Unnatural emotions: Everyday sentiments on a Micronesian Atoll and their challenge to western theory*. Chicago: University of Chicago Press.

Mallon, R. 2007. A field guide to social construction. *Philosophy Compass* 2(1): 93–108.

Mallon, R., and S. Stich. 2000. The odd couple: The compatibility of social construction and evolutionary psychology. *Philosophy of Science* 57: 133–154.

Morsbach, H., and W.J. Tyler. 1986. A Japanese emotion, *Amae*. In *The social construction of emotions*, ed. R. Harré, 289–307. New York: Blackwell.

Popper, K.R. 1972. *Objective knowledge*. Oxford: Oxford University Press.

Rouse, J. 2002. Vampires: Social constructivism, realism, and other philosophical undead. *History and Theory* 41: 60–78.

Rouse, J. 2003. *How scientific practices matter: Reclaiming philosophical naturalism*. Chicago: University of Chicago Press.

Tooby, J., and L. Cosmides. 1992. The psychological foundations of culture. In *The adapted mind: Evolutionary psychology and the generation of culture*, ed. J. Barkow, L. Cosmides, and J. Tooby, 19–136. New York: Oxford University Press.

Wilson, D.S. 2005. Evolutionary social constructivism. In *The literary animal: Evolution and the nature of narrative*, ed. J. Gottschall and D.S. Wilson, 20–37. Evanston: Northwestern University Press.

Scientific Representation, Reflexivity, and the Possibility of Constructive Realism

Tarja Knuuttila

1 Introduction

If there is anything that realists and constructivists, supposed adversaries, appear to agree upon, it is the belief that realism and constructivism are in fact incompatible with one another. Or at least this was the case still a bit more than a decade ago, in the aftermath of the infamous "science wars". One battle line was drawn between realistically inclined philosophers of science and constructivist sociologists of science. For realists the constructivists were more often than not of the "radical social" brand, and for the constructivists the realist philosophers of science provided the target whose unrealistic depictions of science they sought to debunk. Since then, to nearly everyone's relief, things have moved forward both in the field of science and technology studies (hereafter STS) and philosophy of science – and it has begun to seem that they might have more in common than previously believed. On the one hand, as philosophers of science have become increasingly interested in scientific practice, they have become more receptive to some constructivist ideas. On the other hand, the most extreme constructivist tenets have stumbled over their own absurdity, which is what the constructivists soon realized themselves in the discussion on the problem of reflexivity.

In what follows I will delineate some points made in the discussion concerning the problem of reflexivity and show how they are linked to the question of representation. What I aim to argue for is that, ultimately, the constructivist critique of realism relied on the idea of *accurate representation* that they attributed to any realist philosophy of science. This was what they perceived as the root of the "methodological horrors" that, according to Steve Woolgar, haunts scientists'

T. Knuuttila (✉)
Helsinki Collegium for Advanced Studies, University of Helsinki, Fabianinkatu 24 (P.O. Box 4), 00014 University of Helsinki, Helsinki, Finland
e-mail: tarja.knuuttila@helsinki.fi

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_20,

representational endeavours (Woolgar 1988, p. 32). As a consequence, in its wholesale criticism of scientific representation, constructivism actually regenerated the epistemological outlook it sought to avoid. However, if one does not subscribe to the idea of accurate representation much of the constructivist critique of the realist philosophy of science becomes moot. In the constructivist tradition, the radical critique of scientific representation was, indeed, followed by a more moderate investigation into the representational strategies and procedures in scientific practice. These studies, I argue, are reconcilable with and can even contribute to the pragmatic approach to scientific representation. I begin by discussing the problem of reflexivity, after which I move on to the recent philosophical discussion of scientific representation, concluding with a chapter juxtaposing the different versions of constructivism with the different philosophical accounts of scientific representation.

2 The Problem of Reflexivity in Science and Technology Studies

> Within the first few nanoseconds of the relativist big bang, nearly everyone realized that the negative levers were equally applicable to the work of the sociologists and historians themselves (Collins and Yearley 1992, p. 304).

In the 1970s and beginning of the 1980s the newly founded field of science and technology studies engaged in a head-on critique of the "traditional" view of science, typically exemplified by the philosophy of science or functionalist sociology represented by Robert Merton (e.g., the famous norms of science formulated in Merton 1942/1957). The attack proceeded along two main fronts. On the one hand, the adherents of the sociology of scientific knowledge aimed to show that scientific knowledge was shaped by various social factors (e.g., Barnes and Dolby 1970; Mulkay 1976/1991; Bloor 1976/1991). On the other hand, the constructivists went into scientific laboratories to study how the objects of science were constructed (Latour and Woolgar 1979/1986; Knorr-Cetina 1983; Lynch 1985). In both cases they soon encountered the problem of how to justify their own accounts of science given that they aimed to show the interested and constructed nature of the claims of the researchers they studied. In other words, the question became that of what would exempt their own accounts from the same kind of constructivist exposure to which they subjected the scientists' and philosophers' accounts. This problem came to be called the problem of reflexivity in the field of science and technology studies (STS).[1]

[1]To be sure, the STS scholars were not the only ones struggling with the problem of reflexivity at the time. The term was in wide use also, for instance, in sociology and anthropology. In these discussions reflexivity has taken multiple meanings. The same term has been used when talking about modern societies or "modernity", "agents" or subjects, the "participant's" methods

2.1 Reflexivity and the Self-Defeating Character of STS Constructivism

For a scientific study of scientific study the reflexivity of that endeavor should appear obvious. However, the realization of this reflexivity was a matter of more than nanoseconds in STS and yet reflexivity was on the agenda more or less from the beginning of that movement. In David Bloor's influential "The Strong Programme in the Sociology of Knowledge" (Bloor 1976/1991) *reflexivity* was listed among the four tenets that the emerging "social studies of knowledge" should adhere to. The other tenets were *causality*, *impartiality* and *symmetry*. According to Bloor, the sociology of scientific knowledge has to look for the same kind of general causal explanations as other scientific disciplines. Specifically, it should be "concerned with the conditions which bring about belief or states of knowledge" (*id*, p. 7). For Bloor, Barry Barnes and other researchers affiliated with the so-called Edinburgh school this meant finding explanatory factors such as cultural resources, social milieu as well as concerns and interests of different groups. Importantly, the explanations should be impartial with respect to purported truth or falsity of the investigated claims and the same types of cause should be used to explain, symmetrically, both true and false beliefs. Scientific knowledge did not deserve any special treatment and was not to be left to philosophers as putative experts on the rational method.

These kinds of accounts of science were called "interest-explanations" since especially the interests of different groups played a central role in the empirical studies of the Edinburgh group. It is interesting to note, however, that even though reflexivity was on Bloor's list, it did not initially have any prominent place in the emerging sociology of scientific knowledge. It was rather acknowledged as a consequence of the need to seek after general explanations, since in that case the patterns of explanation would have to be applicable, "in principle", also to sociology of scientific knowledge itself.

For the adherents of the "strong programme", the attempt to specify the interests giving rise to scientists' actions meant revealing the social character of scientific knowledge. But if scientific knowledge was regarded as a social product, then what exempted the revealed interests from the same kind of scrutiny? This was asked by Steve Woolgar – who was to become the leading figure of the reflexivists in STS – in his seminal "Interests and Explanation in the Social Study of Science" (Woolgar 1981a). Woolgar urged that the interests should also be investigated instead of being used as unexplicated resources for explanation. In an answer to Woolgar's critique of the "interest explanations", Barry Barnes readily admitted that his conceptions of interests had been "constructed by the analyst so as to perform his explanatory

---

of accounting for their reality and finally about epistemological and methodological issues more generally (e.g., Bourdieu and Wacquant 1992; Knuuttila 2002).

work". What else should they be constructed for, Barnes asked, and added that nothing at all prevents their further study and criticism (Barnes 1981, p. 493).

Woolgar's (1981b) reaction to Barnes's answer is noteworthy, as in retrospect it displays so very clearly how the full-blown STS constructivist agenda begins to take shape. Woolgar namely announces that any attempts to "methodically" arrive at more accurate descriptions of reality are misguided, since descriptions themselves are *constitutive of reality*:

> I'm not saying, then, that the work of MacKenzie and the interest theorists is any more wrong than other attempts at explanation: The artful concealment to which I refer is to be understood as symptomatic of *all* explanatory practice, not as reflection of the motives of particular individuals. So I make no apology for pointing out the significant sense in which *all* such work is *essentially flawed*. The essentially flawed nature of explanation demands our analytic attention and this task should not be set aside in favour of further attempts at explanation (Woolgar 1981b, p. 511).

What Woolgar was in effect suggesting was that instead of explaining what takes place in science, sociologists of science should engage in explicating the various uses of concepts, such as explanation, that they themselves as well as other scientists were using. This project was *ethnomethodological* in nature: the idea was to study the interactional means through which the social order and its products are achieved – or constructed. A good example of this kind of work had already been provided by Michael Mulkay (1976/1991) who had subjected the Mertonian norms to "discourse analytic" scrutiny. From Mulkay's perspective Mertonian norms should not be understood as ideal standards that regulate scientific activity. Instead they are discursive resources that scientists *mobilize strategically* to describe and judge their own and their colleagues' professional behaviour. For Mulkay as well as for many other STS scholars Mertonian norms presented an overly idealized image of science that was being used by scientists to justify the special status of science as an activity not to be interfered with from the outside (see Knuuttila 2012).

While the emphasis was first on the discursive practices, the constructivist programme was soon extended to cover also other aspects of doing science. The turn of the 1980s marks the publication of several important laboratory studies whose method was, to cite one of the protagonists, "direct observation of the *actual site of scientific work* (frequently the scientific laboratory) in order to examine how objects of knowledge are constituted in science" (Knorr-Cetina 1983, p. 117, italics are those of the original; see also Latour and Woolgar 1979/1986; Lynch 1985). The new focus was on how discursive construction was woven together with the experimental, material construction taking place in scientific laboratories. This was especially highlighted by Latour and Woolgar who, in their book *Laboratory life* (1979/1986), asked how the material process taking place in the laboratory was turned into pictures and graphs published in scientific articles. During his 2-year ethnographic study at the Salk institute, Latour observed the transformation processes taking place there: animals were killed, various materials used, the resulting extracts were put into an apparatus, and a sheet of figures was obtained. Now the focus shifted from tubes to figures that were used as input into a computer,

which, in its turn, printed a data sheet. The data sheet was worked on into a curve that was then further processed, redrawn and discussed in the research group, finally ending up as a diagram in a published article. Latour and Woolgar wrote:

> Once the data sheet has been taken to the office for discussion, one can forget the several weeks of work by technicians and the hundreds of dollars, which have gone into its production. After the paper, which incorporates these figures has been written [. . .] it is easy to forget that the construction of the paper depended on material factors [. . .]. Instead, 'ideas', 'theories', and 'reasons' will take their place (Latour and Woolgar 1979/1986, p. 69).

Two observations concerning the STS-style laboratory studies seem especially important. First, the label of social constructivism, not to mention any extreme social constructivism (see Searle 2009) does not apply to them. Construction is understood as a material-cum-social process. In fact, the materiality of this process is often even underlined: the critique is directed against those philosophical and other accounts that tend to forget the material basis of scientific representations. In this the STS constructivists differed for instance from the philosophical constructivists such as Raimo Tuomela, Margaret Gilbert and Michael Bratman, who actually could more properly be called social constructivists than the STS constructivists. The second observation, one that is especially important as regards the question of reflexivity, concerns the method of these studies. Namely, how should one understand the articles produced from the laboratory studies qua representations of how science *actually* works? Was there not something deceptive about this aim given that what these studies set out to argue was precisely the constructed nature of *any* scientific accounts. Could such ethnographic studies, or texts reporting them, really succeed in avoiding the conventional explanatory schemes themselves, and thus circumvent suspicions about their own constructed nature?

Indeed, the STS reflexivists, with Steve Woolgar and Malcolm Ashmore in the lead, soon launched a forceful critique of the ethnomethodological pretension of being in the possession of a privileged method with which one could somehow find out "what actually goes on in science". It is telling that Latour and Woolgar added to the second 1986 edition of their *Laboratory life* a postscript that discussed the ethnographic method and reflexivity. Also the word "social" was dropped from the subtitle of the book that now became: "The construction of scientific facts". This is no surprise as Woolgar had by then, as pointed out above, already engaged in the critique of also laboratory studies, attacking their instrumental conception of ethnography. Such a conception applies relativist epistemology only *selectively* – to other scientists' accounts – whereas one's own accounts are presented realistically. An instrumental ethnographer, according to Woolgar (1982, p. 485), tends to be after news, of "finding things to be other than you supposed they were". In this case, the news was, more often than not, that scientific facts are constructed and that science does not differ from non-science – it is as "social", "contingent", "local", "situated", and so forth, as any other social activity.

In and of itself the newly found social character of science was not very stunning news. To render it such, an alternative "traditional" conception had to be put up, with which the novel constructivist conception of science could be contrasted. This

strategy provided, for quite a long period, a popular way to open up a STS-article (which is still in use, actually). The alternative, old-fashioned and even damaging view, was provided, more often than not, by the "philosophical version" of science as rational activity oriented at finding the truth. Typically, not much was said about this view – it was merely alluded to, or presented briefly, in an uncontextual and general manner. The impression a reader easily gets in browsing through past and even some present STS literature is that the "traditional view" invoked is mostly a rhetorical construct, the rationale of which is to underline the novelty and epoch-making character of the (constructivist) views professed.

The irony of this situation is that the constructivist is not herself practising what she is preaching. As it was put by Woolgar (1983, p. 245), the preferred constructivist position on reality and its representation is the mediative one according to which "there is nothing inherent in the character of real world objects which uniquely determines the accounts of those objects". This conviction is, then, seasoned with differing amounts (depending on the constructivist in question) of constitutive intuitions that we construct realities by way of accounting them. Yet, when it comes to contrasting the descriptions of the constructivist ethnographer and those of the scientists she studies (or philosophers' and other traditionalists' accounts), there seems to be no doubt about *whose* story is supposed to *fit* reality best. In other words, according to constructivists, scientific representations should not be regarded as truthful representations of real objects or processes: there is no determinable correspondence between our scientific representations and the reality they aim to explain or describe. Yet in arguing for this view, the constructivist has created one more representation, but now *this* representation is offered as a truthful depiction of its object, that is science, as it "actually" happens. The question, then, is how to meet the reflexive challenge?

Although the reflexivists' critique of the paradoxes of wholesale constructivism was perceptive, their proposed solution to the problem was less so.[2] Since it seemed to them that there was no adequate solution to the problem of reflexivity, they suggested that instead of trying to solve the problem it should rather be "celebrated". It was proposed that with different textual methods the "monster" of reflexivity could be "simultaneously kept at bay and allowed a position at the heart of our enterprise" (Woolgar 1982, p. 489). Such methods were, for instance, the "second voice device" (Woolgar and Ashmore 1988), and other kinds of dialogues

[2]To be sure, this reflexive problem that the constructivists faced was but one version of the problem of relativism. See Mary Hesse (1980) for a more philosophically interesting defense of STS constructivism. She claimed that a relativist means different things than an objectivist by such expressions as "truth", "knowledge" and "grounds". A "truth" for a relativist, for example, means that which meets the criteria of truth in a local culture. Therefore the non-relativist attempt to show that the relativists' claims are self-defeating does not succeed, since it makes use of the senses of "true" and "knowledge" excluded by the relativists' claims. But then a new problem appears: it is not clear that the relativist and her critic are any longer engaged in the same philosophical controversy (see Tollefsen 1987, p. 211).

trying to display their constructed nature. Their aim was to shatter the reader's supposed "naive belief" in the text and make her aware of the text's artificial nature by constructing it so that it more or less deconstructs itself. As a result the text becomes more of an epistemological project directed to question our alleged epistemological habits (i.e. naïve realism) than any scientific representation of the empirical subjects studied. According to Woolgar (1982, p. 492) "reflexive ethnography need not entirely exclude the production of news about laboratories; this becomes an incidental product of research, rather than its main objective."[3]

No wonder, then, that the reflexivist programme never took off in science and technology studies: readers were still more interested in the news about laboratories. Indeed, one big problem of the STS reflexivism was its disregard of the reception, in other words, the readers of texts. Especially when it comes to articles published in scientific journals, the scientists reading them are certainly not naïve realists, unaware of the laborious experimental and theoretical processes lying behind the published results. It is as if the reflexivists fell prey to their own trap in fixing their gaze on the *finished* product, the text. Being constructivists, one would have expected them to pay more attention to the processes in which scientific representations were produced and used. And this was precisely the direction to which they turned after less than a decade of reflexive "wrighting" (Ashmore 1989).[4]

2.2 The Practices of Scientific Representation

In retrospect the collection *Representation in Scientific Practice* edited by Lynch and Woolgar (1990) paved the way for how constructivist STS has approached scientific representation ever since.[5] Several contributions of the book meticulously

[3]This proposal was clearly a part of a larger current sweeping over humanities and social sciences in the 1970s and early 1980s. In the field of historiography, Hayden White (e.g., 1973) urged historians to pay attention to the historical narratives themselves, to their fictional and artificial nature. Clifford Geerz (e.g., 1973), in turn propagated for the same kind of programme in anthropology. Spencer notes how, as a result of Geerz's emphasis on writing, his hermeneutic approach "tries above all to close the hermeneutic cycle by limiting his readers' access to that which he wants to interpret himself" (1989, p. 149).

[4]The reflexivist agenda did not disappear from STS, or from sociology in general, instead it was reformulated. Consider for instance how Woodhouse et al. (2002, p. 307) advocate activism: "[I]n as much as there always are more research questions than time to study them, it seems hard to miss the possibility of extending the individual-level reflexivity of the 1980s to the field more generally: what social processes are setting our collective agendas; is the agenda-setting process a laudable one [. . .]".

[5]The book *Representation in Scientific Practice Revisited* (Coopmans et al. 2014) takes its inspiration from its predecessor, taking into account novel forms of image production, especially such as digital image processing and new kinds of tactile and haptic representations.

follow the "assembly line", the processes of constructing scientific representations. From this point of view scientific representation appears as a subtle "dialectic of gain and loss" (Latour 1995). It is not just a question of reduction or simplifying. Some methods of representation further fragment, upgrade or define the specimen in order to reveal its details, whereas others add visual features for the purposes of clarifying, extending, identifying, etc. Often the aim of scientific representation is to mould the scientific object so that it can assume a mathematically analysable form or to be more easily described and displayed by using different textual devices (see also Latour 1990; Lynch 1985, 1990). Scientific representation widens in these studies into an expanded process of circulating and arranging diverse pictures, extracts, "tissue cultures", photographic traces, diagrams, chart recordings and verbal accounts. Representations become things that are worked upon, being ultimately "rich depositories of 'social' actions" as Lynch and Woolgar sum up the approach in their introduction to the volume (1990, p. 5).

Any ambitious epistemological programmes are largely left behind in the book, although some authors of the book still aim to "explode" the supposed homogeneous conception of representation in order to make room for the "deeds performed, when those [representational] items are embedded in action" (Lynch 1994, p. 146). However, it seems that this constructivist agenda is rather traditional when it comes to its approach on scientific representation. Firstly, in an effort to deconstruct the notion of representation the protagonists rely on what is commonly taken as representation already in their choice of case studies. Thus rather than exploding the notion of representation, these cases actually reveal instead what a complicated phenomenon scientific representation is. Secondly, these studies still seem to be engaged in showing us what *really* goes on in scientific representation. That is, the reflexivist worries are clearly ignored. Thirdly, although these studies focus on scientific representations and not on the phenomena they are supposed to be representing, they give us clues as to how, through the laborious art of representing, scientists are seeking and gaining new knowledge.

What is essential for this constructivist perspective is the focus on what one does with scientific representations and how various representational media are utilized in the process of scientific representation. The artificial features of scientific representations do not render them as contingent social constructions, but rather result from well-motivated epistemic strategies that in fact enable scientists to know more about their objects. Thus what is not threatened is the possibility to represent the external world and to gain knowledge from it through representation; what is questioned, however, is the idea that scientific representations are some kind of transparent imprints of reality with a single determinable relationship to their targets. Now the question is, what could possibly be so threatening about *this* view of scientific representation? Does it make constructivism and scientific realism incompatible with each other? The answer, I suggest, depends on how we conceive of scientific representation. With this in mind let me consider next the current philosophical discussion on scientific representation.

3 From Semantic to Pragmatic Accounts of Representation

The question of representation arose in the philosophy of science only relatively recently, although the idea of representing the world accurately has been central to our common conception of science and to the philosophical discussion of realism (e.g., Godfrey-Smith 2003, pp. 176–177). Yet it was not until the beginning of the 2000s that representation as a specific topic began to interest philosophers of science more generally. Once started, the philosophical discussion focused almost exclusively on scientific representation in the context of modelling. The discussion was largely motivated by the supposition that models give us knowledge because they represent their supposed real-world target systems more or less faithfully, in relevant respects and to a sufficient degree. Such an assumption already suggests that there is a special sort of relationship between a model and its target, and the question became one of how to analyse such a relationship. Could representation be analysed in such terms as isomorphism or similarity, or is something else needed to establish the representational relationship? To this question philosophers of science have given various answers, which have far-reaching implications for how the epistemic value of models is to be understood.

The conviction that representation can be accounted for by solely reverting to the properties of the model and its target system is part and parcel of the semantic/structuralist approach to scientific modelling. According to the semantic conception, models specify structures that are posited as possible representations of either the observable phenomena or, even more ambitiously, the underlying structures of real-world phenomena. The semantic/structuralist conception of scientific representation was originally cast in terms of isomorphism: a given structure represents its (real-world) target system if they are structurally isomorphic to each other (e.g., van Fraassen 1980; Suppe 1974, 1989). Isomorphism refers to a kind of mapping that can be established between two structures and preserves the relations among the elements. Giere (1988) in turn suggested similarity as the basis of the representational relationship in his reformulation of the semantic approach, but he has later come to think that similarity fits better the pragmatic approach to scientific representation (see below).

The recent philosophical discussion has found the analysis of representation in terms of isomorphism lacking in many respects. Firstly, isomorphism does not have the right formal properties to capture the nature of the representational relationship: it is a symmetric, transitive and reflexive relationship whereas representation is not.[6] Secondly, it does not leave room for misrepresentation. The idea that representation is either an accurate depiction of its object or not a representation at all does not

[6]These points derive from Nelson Goodman's famous critique of similarity (Goodman 1968). For reasons of space I cannot deal with them in detail, and readers are referred to Suárez (2003) and Frigg (2006). Suárez has also directed this line of critique towards the similarity account, but the philosophers of science currently favoring a looser (i.e. not mathematical) notion of similarity all tend to take into account users and use (e.g., Giere 2004, 2010, see above).

fit actual representational practises. Thirdly, structure sharing is not necessary for representation. Scientific practise is full of examples of inaccurate models, which are difficult to render as isomorphic with their targets. Fourthly and perhaps most importantly, isomorphism does not capture the directionality of representation. We usually want the model to represent its target but not vice versa.

Structuralists have attempted to counter these criticisms in two ways, either amending the structural account in adding directionality to it (e.g., Bartels 2006), or trying to weaken the conditions that isomorphism imposes on representation by suggesting different morphisms such as homomorphism (Lloyd 1988; Bartels 2006; Ambrosio 2007) or partial isomorphism (Bueno 1997; French and Ladyman 1999; da Costa and French 2003). Both of these notions attempt to do away with the problems of misrepresentation and non-necessity. It is worthy of note that in defending homomorphism as an alternative to isomorphism Bartels (2006) suggests that it has to be complemented with a representational mechanism connecting the representational vehicle to its target. Such mechanism would capture the directionality of representation. This seems to suggest that, whereas Bartels makes an effort to give a fully-fledged analysis of representation, it is questionable whether other structuralists ever attempted to present any necessary and sufficient conditions of scientific representation. Indeed, in a footnote French and Ladyman (1999, p. 119) clarify that they "are not claiming that the gap between a theory or model and reality can be closed simply by a formal relation between model-theoretic structures". Yet it seems that in their conviction that "it involves isomorphism" (French 2003) the structuralists have usually left the rest unexamined. Consequently structural relations seem to provide the *privileged foundation* on which our knowledge rests. In this context it is interesting to note that the (in)famous attack on representation by Richard Rorty (1980), is actually an attack on "privileged representations" that, according to Brandom, are supposed to possess "a natural or intrinsic epistemic privilege, so that their mere occurrence entails that we know or understand something. They are self-intimating representing: having them counts as knowing something" (2009, p. 6).

Many of the problems in the semantic/structuralist account of representation are directly related to the fact that scientific representation is a relation between a representational vehicle (e.g., a model) and a real target, and thus a mere mathematical relation between two structures fails to capture some of its inherent features – and makes too stringent demands on actual scientific representations. According to the pragmatists these problems will be cured if it is recognized that representation cannot be based only on the respective properties of the representational vehicle and its target system. Intended uses or users' intentions, both create the directionality needed to establish a representational relationship and introduce the necessary indeterminateness into it (given that human beings as representers are fallible). However, this comes at a price. When representation is grounded primarily on the specific goals and representing activity of humans as opposed to the properties of the representative vehicle and the target object, it is deprived of much of its explanatory content: not much insight into the epistemic value of modelling is gained in claiming that models give us knowledge *because* they are *used* to represent their target objects.

One strategy to deal with this problem is to add to one's account of representation a further stipulation concerning its success. Rather unsurprisingly, then, what has earlier been presented as an analysis of the representational relationship, i.e., isomorphism (van Fraassen 2008) or similarity (Giere 2010), is now suggested as a success criterion.[7] As for isomorphism, it poses too stringent a condition on the success of representation in the light of scientific practise. The case of similarity is trickier. On the one hand, it does not really supply any user-independent success criterion in that it is the users who identify the "relevant respects and sufficient degrees" of similarity. Giere admits this, arguing that an *agent-based* approach "legitimates using similarity as the basic relationship between models and the world" (2010, p. 269).

Another possibility for a pragmatist is to go deflationary all the way, as Suárez (2004, 2010) has done, and resist saying anything substantive about the representational relationship or its success, in other words whether they rest on isomorphism, similarity or denotation, for instance. According to Suárez, substantive accounts of representation err in trying to seek for some deeper constituent relation between the source and the target which could then, as a by-product, explain why the source is capable of leading a competent user to consideration of a target, and why scientific representation is able to sustain "surrogate reasoning". Hence he explicitly denies any privileged relationship between a representational vehicle and its target.[8] Instead, Suárez builds his analysis directly on the aforementioned by-products. His inferential account of scientific representation is two-sided, consisting of *representational force* and the *inferential capacities* of the representational vehicle. Representational force results from the practice of using a particular representational vehicle as a representation, determining its intended target. In addition to that the vehicle must have *inferential capacities* that enable the informed and competent user to draw valid inferences regarding the target. The success of representation also implies that there are some norms of inference in place distinguishing correctly drawn inferences from those that are not (Suárez 2010). These features of Suárez's proposal in particular, and those of pragmatic accounts of representation in general, make them interesting from the perspective of the possible reconciliation of constructivism and realism.

[7]Looking at the structuralist-pragmatist controversy from this perspective suggests that at least partly the two adversaries are arguing at cross-purposes. The structuralists seem to have been more interested in the question of what would justify a representational relationship whereas the pragmatists have focused on the use of scientific representations. Chakravartty (2010) has attempted to capture this contrast with his distinction between informational versus functional theories of scientific representation.

[8]This has recently been argued also by Weisberg (2007), according to whom "[m]odels do not have a single, automatically determinable relationship to the world" (p. 218).

4 The Possibility of Constructive Realism

As discussed above, one strand of the constructivism versus realism debate has revolved around the question of representation. What constructivists in the field of science and technology studies have typically wanted to contest is the idea of science providing us accurate representations of the world. This critique, however, soon turned against the STS constructivists themselves in the form of the problem of reflexivity, as they nevertheless wanted show *how* scientific representations, facts and objects were constructed in *actual* scientific practice. Thus while the STS scholars aimed to expose others' scientific representations as interested and constructed, they simultaneously treated their own accounts as realistic depictions of science "as it happens". This problem was soon noticed without any viable answer given to it, and eventually the STS constructivists turned to studying the practices of representation instead of worrying about the relationship of scientific representations to their supposed real-world targets. However, it seems fair to say that the underlying motivation of contesting the idea of scientific representations as accurate depictions of real-world phenomena remained the same also in the new constructivist programme of concentrating on representation in practise.

As for the philosophical discussion of scientific representation, one central disagreement has precisely concerned the question of whether scientific representation should be understood in terms of accuracy or not. But what is meant by accuracy? Nothing even remotely like mirroring or copying seems to work when it comes to scientific representation. Accuracy in this context is captured by the idea of structure sharing: "isomorphism provides us with a criterion for what counts as accurate" (Frigg 2006, p. 12). Consequently, a scientific representation is accurate if it depicts a structure that is isomorphic with the underlying structure of its real-world target system. Clearly, also the other kinds of structuralist accounts based on various types of morphisms still attempt to latch onto this kind of accuracy.

From the perspective of the constructivist epistemological programme it seems clear that the semantic/structuralist account provides precisely the kind of philosophical view that it seeks to challenge. This becomes apparent once we consider some central features of the semantic/structuralist view. The first thing to note about it is that it provides a *strong realist account* of representation that simultaneously gives an analysis of scientific representation and an objective criterion for its accuracy.[9] Secondly, it provides us a dyadic analysis of representation that (largely) reduces the relationship of representation to the properties of representational vehicles and their targets. Consequently there is neither room, nor function for users and the various intended uses of models. In other words, although no proponent of the semantic/structuralist view would surely deny that models are being used in actual scientific practise, these social aspects of representing are not deemed

[9]This need not be the case, however. One might also pursue an empiricist argument as van Fraassen (1980) does. For him the isomorphism at stake concerns the relationship between the model and the structures of (empirical) "appearances".

crucial when it comes to accounting for the *epistemic* value of models. Thirdly, the attempt to reduce the relationship of representation to isomorphism extracts from the actual scientific models a privileged layer, the structure, in virtue of which accurate representation is possible. What this amounts to is the claim that "the specific material of the models is irrelevant; rather it is the structural representation, in two or three dimensions, which is important" (French and Ladyman 1999, p. 109).

Consequently, the structuralist/semantic account provides a strong realist account of representation that has no important place for either the social aspects of representing, or the various characteristics of the actual representational media used – that is whether the model is expressed, for example, with symbolic, iconic, or diagrammatic means, or as a 3D physical object. This kind of conception of scientific representation is clearly incompatible with constructivism. Yet, it provides just one philosophical approach to scientific representation – which fact seems to have escaped the constructivists' notice. Thus it seems fair to conclude that in their wholesale attack on representation the reflexivists (and some representation-hostile neo-pragmatists like Rorty) were themselves relying on an unduly stringent notion of representation.

The pragmatic approaches to representation are, in contrast, inherently triadic. This means that the users and the purposes, for which scientific representations are constructed, cannot be neglected in accounting for the representational relationship and its success. It also implies that the basic unit of analysis as regards scientific representation cannot be restricted to the model-target dyad (e.g., Knuuttila 2011) that opens up space for considering the social aspects of representation and the process of constructing and using representations. Although the pragmatic accounts of scientific representation have not so far targeted the epistemic value of the *process of* constructing representations, several accounts of model construction have been recently presented (e.g., Peschard 2011; Knuuttila and Boon 2011). Apart from the epistemic value inherent in the process of constructing and using representations, another aspect of scientific representation stressed by the constructivist studies is due to the way scientists utilize the specific characteristics of various representational media. This line of investigation has only just begun in the philosophy of science (e.g., Gelfert 2011; Knuuttila 2011; Vorms 2012; Chandrasekharan and Nersessian forthcoming). It is however implied by the deflationary account of scientific representation by Suárez (2004, 2010) that stresses the importance of the inferences that scientific representations license. For these inferences, the representational media used clearly matters.[10]

To conclude, there is a substantial overlap and hence a possibility for a fruitful dialogue between the pragmatic understanding of scientific representation and the constructivist studies of the representational practises in science. Both approaches are also compatible with moderate constructive or perspectival realism that is purpose-oriented, inter-subjective, and instrument-using in character (Giere 1988,

[10]The cognitive importance of the specific mode of a representation is already a well-researched topic in cognitive science (see e.g., Zhang 1997; Hutchins 1995).

2006). Such constructive realism does not expect our representations to be accurate depictions of their target systems, yet it approaches them as important mediators of knowledge about the real world. Last but not least, from the perspective of better understanding representation in scientific practise, the deflationary nature of the pragmatic account need not be regarded as its weakness. Quite the contrary, it might prompt philosophers of science to engage in collaboration with scientists from neighbouring disciplines, such as cognitive science and semiotics, to provide more empirical flesh to it.

References

Ambrosio, C. 2007. *Iconicity and network thinking in Picasso's Guernica: A study of creativity across the boundaries.* Unpublished Ph.D. thesis, University College London.

Ashmore, M. 1989. *The reflexive thesis: Wrighting sociology of scientific knowledge*. Chicago: Chicago University Press.

Barnes, B.S. 1981. On the 'hows' and 'whys' of cultural change (response to Woolgar). *Social Studies of Science* 11: 481–498.

Barnes, B.S., and R.G.A. Dolby. 1970. The scientific ethos: A deviant viewpoint. *Archives Européennes de Sociologie* 11: 3–25.

Bartels, A. 2006. Defending the structural concept of representation. *Theoria* 55: 7–19.

Bloor, D. 1976/1991. *Knowledge and social imagery*, 2nd ed. Chicago: University of Chicago Press.

Bourdieu, P., and L.J.D. Wacquant. 1992. *An invitation to reflexive sociology*. Cambridge: Polity Press.

Brandom, R. 2009. Global anti-representationalism? http://www.pitt.edu/~brandom/index.html. Accessed 24 Nov 2009.

Bueno, O. 1997. Empirical adequacy: A partial structure approach. *Studies in the History and Philosophy of Science* 28: 585–610.

Chakravartty, A. 2010. Informational versus functional theories of scientific representation. *Synthese* 172: 197–213.

Chandrasekharan, S., and N.J. Nersessian. forthcoming. Building correspondence: How the process of constructing models leads to discoveries and transfer in the engineering sciences. *Erkenntnis*.

Collins, H., and S. Yearley. 1992. Epistemological chicken. In *Science as practice and culture*, ed. A. Pickering, 301–326. Chicago: University of Chicago Press.

Coopmans, C., J. Vertesi, M.E. Lynch, and S. Woolgar (eds.). 2014. *Representation in scientific practice revisited*. Cambridge, MA: The MIT Press.

da Costa, N.C.A., and S. French. 2003. *Science and partial truth. A unitary approach to models and scientific reasoning*. New York: Oxford University Press.

French, S. 2003. A model-theoretic account of representation (or, I don't know much about art . . . but I know it involves isomorphism). *Philosophy of Science* (Proceedings) 70: 1472–1483.

French, S., and J. Ladyman. 1999. Reinflating the semantic approach. *International Studies in the Philosophy of Science* 13: 103–121.

Frigg, R. 2006. Scientific representation and the semantic view of theories. *PhilSci-Archive*. http://philsci-archive.pitt.edu/2926/. Accessed 18 Apr 2013.

Geertz, C. 1973. *Interpretation of cultures*. New York: Basic Books.

Gelfert, A. 2011. Mathematical formalisms in scientific practice: From denotation to model-based representation. *Studies in the History and Philosophy of Science* 42(4): 272–286.

Giere, R.N. 1988. *Explaining science: A cognitive approach*. Chicago/London: University of Chicago Press.
Giere, R.N. 2004. How models are used to represent reality. *Philosophy of Science* (Symposia) 71: 742–752.
Giere, R.N. 2006. *Scientific perspectivism*. Chicago: University of Chicago Press.
Giere, R.N. 2010. An agent-based conception of models and scientific representation. *Synthese* 172: 269–281.
Godfrey-Smith, P. 2003. *Theory and reality: An introduction to the philosophy of science*. Chicago: University of Chicago Press.
Goodman, N. 1968. *Languages of art: An approach to theory of symbols*. Indianapolis: Bobbs-Merrill.
Hesse, M. 1980. *Revolutions and reconstructions in the philosophy of science*. Brighton: Pergamon Press.
Hutchins, E. 1995. *Cognition in the wild*. Cambridge, MA: The MIT Press.
Knorr-Cetina, K.D. 1983. The ethnographic study of scientific work: Towards a constructivist interpretation of science. In *Science observed: Perspectives on the social study of science*, ed. K.D. Knorr-Cetina and M. Mulkay, 1–18. London: Sage.
Knuuttila, T. 2002. Signing for reflexivity: Constructionist rhetorics and its reflexive critique in science and technology studies [52 paragraphs]. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 3(3). http://www.qualitative-research.net/index.php/fqs/article/view/828/1798. Accessed 30 Mar 2013.
Knuuttila, T. 2011. Modeling and representing: An artefactual approach. *Studies in the History and Philosophy of Science* 42: 262–271.
Knuuttila, T. 2012. Contradictions of commercialization: Revealing the norms of science? *Philosophy of Science* 79: 833–844.
Knuuttila, T., and M. Boon. 2011. How do models give us knowledge? The case of Carnot's ideal heat engine. *European Journal for Philosophy of Science* 1(3): 309–334.
Latour, B. 1990. Drawing things together. In *Representation in scientific practice*, ed. M. Lynch and S. Woolgar, 19–68. Cambridge, MA: The MIT Press.
Latour, B. 1995. The 'Pédofil' of Boa vista: A photo-philosophical montage. *Common Knowledge* 4(1): 144–187.
Latour, B., and S. Woolgar. 1979/1986. *Laboratory life: The construction of scientific facts*, 2nd ed. Princeton: Princeton University Press.
Lloyd, E. 1988. *The structure and confirmation of evolutionary theory*. Princeton: Princeton University Press.
Lynch, M. 1985. *Art and artifact in laboratory science: A study of shop work and shop talk in a research laboratory*. London: Routledge and Kegan Paul.
Lynch, M. 1990. The externalized retina: Selection and mathematization in the visual documentation of objects in life sciences. In *Representation in scientific practice*, ed. M. Lynch and S. Woolgar, 153–186. Cambridge, MA: The MIT Press.
Lynch, M. 1994. Representation is overrated: Some critical remarks about the use of the concept of representation in science studies. *Configurations: A Journal of Literature, Science and Technology* 2: 137–149.
Lynch, M., and S. Woolgar (eds.). 1990. *Representation in scientific practice*. Cambridge, MA/London: The MIT Press.
Merton, R.K. 1942/1957. Science and democratic social structure. In *Social theory and social structure. Revised and enlarged edition*, ed. R.K. Merton, 550–561. Reprinted in Glencoe: The Free Press.
Mulkay, M. 1976/1991. Norms and ideology. In *Sociology of science. A sociological pilgrimage*, ed. M.J. Mulkay, 62–78. Milton Keynes: Open University Press.
Peschard, I. 2011. Making sense of modeling: Beyond representation. *European Journal for Philosophy of Science* 1: 335–352.
Rorty, R. 1980. *Philosophy and the mirror of nature*. Oxford: Basil Blackwell.

Searle, J.R. 2009. Why should you believe it? *New York Review of Books*, September 24, LVI, 14: 92.
Spencer, J. 1989. Anthropology as a kind of writing. *Man (N.S.)* 24: 145–164.
Suárez, M. 2003. Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science* 17: 225–244.
Suárez, M. 2004. An inferential conception of scientific representation. *Philosophy of Science* (Symposia) 71: 767–779.
Suárez, M. 2010. Scientific representation. *Blackwell's Philosophy Compass* 5(1): 91–101.
Suppe, F. 1974. *The structure of scientific theories*. Urbana: University of Illinois Press.
Suppe, F. 1989. *The semantic conception of theories and scientific realism*. Urbana/Chicago: University of Illinois Press.
Tollefsen, O. 1987. The equivocation defense of cognitive relativism. In *Self-reference: Reflections on reflexivity*, ed. S.J. Bartlett and P. Suber, 209–217. Dordrecht: Martinus Nijhoff.
van Fraassen, B. 1980. *The scientific image*. Oxford: Oxford University Press.
van Fraassen, B. 2008. *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.
Vorms, M. 2012. Formats of representation in scientific theorizing. In *Representations, models, and simulations*, ed. P. Humphreys and C. Imbert, 250–273. London: Routledge.
Weisberg, M. 2007. Who is a modeler? *British Journal for the Philosophy of Science* 58: 207–233.
White, H. 1973. *Metahistory: The historical imagination in nineteenth-century Europe*. Baltimore: The Johns Hopkins University Press.
Woodhouse, E., D. Hess, S. Breyman, and B. Martin. 2002. Science studies and activism: Possibilities and problems for reconstructivist agendas. *Social Studies of Science* 32(2): 297–319.
Woolgar, S. 1981a. Interests and explanation in the social study of science. *Social Studies of Science* 11: 365–394.
Woolgar, S. 1981b. Critique and criticism: Two readings of ethnomethodology. *Social Studies of Science* 11: 504–514.
Woolgar, S. 1982. Laboratory studies: A comment on the state of art. *Social Studies of Science* 12: 481–498.
Woolgar, S. 1983. Irony in the social study of science. In *Science observed: Perspectives on the social study of science*, ed. K.D. Knorr-Cetina and M. Mulkay, 239–266. London: Sage.
Woolgar, S. 1988. *Science: The very idea*. London: Tavistock.
Woolgar, S., and M. Ashmore. 1988. The next step: An introduction to the reflexive project. In *Knowledge and reflexivity: New frontiers in the sociology of knowledge*, ed. S. Woolgar, 1–11. London: Sage.
Zhang, J. 1997. The nature of external representations in problem solving. *Cognitive Science* 21: 179–217.

The Limits of Realism in the Philosophy of Social Science

David-Hillel Ruben

There is an old Russian proverb, quoted in Vladimir Medem's autobiography, that says: "an individual in Russia was composed of three parts: a body, a soul, and a passport". It isn't only that there are these three aspects of a person, but moreover that somehow the three are connected. How so?

Just as most philosophers (excluding eliminativists and those who posit pre-established harmony, for example) believe that there are both the physical and mental realms and that they must be connected in some way-the options range from identity through supervenience to causation-so too it is plausible to believe that the social and the mental (and from now on, we can add "the physical" along with the mental, without always further specifying it explicitly) have some sort of important relationship, whatever it might be. I speak about the mental; the literature often speaks of the individual. For our purposes, these come to the same, since the mental facts in which we are interested are mental facts about individual persons. The term "individual" does not capture the physical facts that may have to be added to the mental in any event, to account for the social, since the relevant physical facts might not even be about persons at all.

Where shall we place human action in Medem's tripartite division? There is some unclarity about this in the literature of (so-called) methodological individualism. Human action itself divides into social and non-social action. It would be an unnecessary digression to make this distinction explicit, but intuitively voting, cashing cheques, and engaging in a ritual are social actions; climbing a mountain, riding a horse, and building a shelter are not. There will be many cases whose classification will be undecidable without a precise and explicit account of this distinction, but other cases, like the ones mentioned above, will be clear. Facts about social actions are part of the social; facts about non-social actions are to be included within the mental.

D.-H. Ruben (✉)
Department of Philosophy, Birkbeck University of London, 12-14 Gower Street, London WC1E 6DP, UK
e-mail: d.ruben@bbk.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_21,

Here is a compelling if somewhat minimalist thought: if there were no beings with a mental life, there could be no social world. Is there more we can say about the tripartite relationship Medem mentions, of the mental and physical on the one hand and the social on the other?

Let's assume for the sake of argument that reductive identity (between the mental and the physical, between the social and the mental) is not on the table. As Ned Block says, "For nearly thirty years, there has been a consensus . . . that reductionism is a mistake . . . " (and that was said already 16 years ago!) (Block 1997). If reductionism is a mistake, it is not only identity (between the mental and the physical or between the social and the mental) that is in trouble; nomological equivalence would be as well, since most accounts of reduction require one or the other of these ideas.

The reductive identity being dismissed by Block and others is ontological, not just discourse or theory reduction. Of course, not all cases of failure of reduction of one discourse to another entail failure of reduction in an ontological sense. For example, the inability to "translate" physical object discourse into sense data discourse ("without remainder") does not on its own show that physical objects are not just sets of sense data. But if reductive identity fails in the mental-physical or the social-mental cases because of failure of identity or nomic equivalence between properties or entities in the allegedly reducing and to-be-reduced discourses, this will certainly have ontological implication. It would show that we have two distinct or non-identical sets of properties or particulars, however they might otherwise be related.

Some of the reasons for this failure in the social-to-mental case can be learned from the failure of reduction of the mental to the physical. Just to summarise a very long story, type reduction fails because of multiple realisability (and the unacceptability of infinite disjunctions or of heterogeneous disjunctions for reductive purposes). Token reduction fails because of its dependence on type reduction (there are other reasons for failure of token reduction but again the story is too long to repeat here).

There is an added problem with regard to the purported reduction of the social: a circularity argument. Any specification of the allegedly reducing base will certainly make use of social concepts. So conceptual reduction of the social to the non-social fails. But what of the ontological issue rather than the conceptual one? In my view, the allegedly reducing base will also contain references to individual social entities of some sort, and will also require the existence of social properties, and not just make use of social concepts, so allegedly ontological reduction, and not only conceptual reduction, will fail the circularity test.[1]

What are the other options for understanding the social/non-social relation? In a way similar to the dialectic in the philosophy of mind, supervenience, it was once hoped, would offer an alternative to reduction that at one and the same time would be both non-reductive but also non-dualistic. On this approach, just as the mind and

[1] I have given several of these arguments in Ruben (1984).

body could be tied by supervenience sufficiently tightly to avoid dualism but not so tightly to collapse them into one, so too for the social and the mental: the hope was that the social could preserve its integrity but without claiming true autonomy or independence. Supervenience must provide the best example ever of both having and eating the proverbial cake.

One of the earliest examples of this hope in the philosophy of social science was Greg Currie's "Individualism and Global Supervenience" (Currie 1984). For reasons that are especially appropriate in the social world, any type of supervenience that is likely to hit the mark will be global rather than individual or even regional, because the way in which things are socially might depend on the way in which things are with individuals far removed in space and time from the particulars that constitute the social fact under consideration or from the region they inhabit. La Guardia may have been the mayor of New York, but in order for that to be true, an indeterminate number of other people had to hold certain desires and beliefs. For some practice a person engages in, in order for it to count as a tradition, there must be persons engaging in a similar practice many times in the past. Social facts require many individual mental facts spread in time and space whose extent is hard to circumscribe in advance.

As Currie says, "... it is the totality of individual facts which determines the totality of social facts" (Currie 1984, p. 345). The supervenience of the social on facts about individuals, for him, comes to this: consider two possible worlds, u and w, which have identical individual (mental and physical) histories up to and including time t. If so, then the same social facts are true in both u and w at t. Any two worlds that are indiscernible in terms of individual (physical and mental) facts up to and including t are indiscernible with respect to social facts at t; no social difference without an individual or physical difference. Note that the converse is not the case: the individual facts won't also supervene on the social.[2] There can certainly be socially indistinguishable worlds that differ in terms of individual facts. Indeed, multiple realisability insures that this is so.

Let us put it this way: that something has a social property (e.g., being a pound note or dollar bill) supervenes on agents believing that some physical thing has that, or other, social property(−ies). Once God fixes the way in which the mental life of agents is in some world and the contents of their beliefs and desires, and He also fixes the way the social world is in that same world, then if He brings into being another world with its agents having the same mental life as in the first, there is nothing more that He needs to decide about what the social world in that second world will be like. It must be the same as it was in the first world.

The use of supervenience by philosophers of social science has a long history. Currie's article was written in 1984. Christian List and Philip Pettit's *Group Agency* (2011) uses the same idea. Philosophers of social science many times in between have invoked it.

[2]I count identity as the strongest form of supervenience and identity is, of course, symmetric. The asymmetry claim holds only for cases of supervenience other than identity.

List and Pettit explicitly assert that supervenience of the social on the mental allows them to steer a middle path between the two traditions of reduction and eliminativism on the one hand and emergentism on the other, both of which they reject. It is true that the autonomy they claim for the social is epistemological rather than ontological, but it is hard to see why this should be so, given their other views. Supervenience is an alleged ontological relation and their own discussion is conducted in ontological terms: one set of facts determines or fixes another set (e.g., p. 65). If supervenience works in the way they want, it ought to yield some sort of ontological and not just epistemological autonomy.

Suppose you find the idea of the supervenience of the social on the (individual) mental a plausible view. Supervenience is an objective (in one sense of that very slippery term) relation that does not depend on thinking, wanting or willing it to be so, or on anything human agents do. It is true that in the case of the alleged supervenience of the social on the mental, the supervenience base includes individuals' beliefs, desires, wishes, non-social actions, and any other mental item one might regard as important for inclusion, but once the supervenience base is specified, the relation between the supervenience base and what supervenes on it is objective, in the sense that that relation holds, if it does, regardless of anything agents might think or do.

An analogy might be with theories of political obligation. If one looks at a traditional discussion of the grounds of political obligation, the answers given to the question of the grounds of political obligation break down in to two main categories. For example, three of the five theories that D.D. Raphael discusses, social contract, consent, and the general will, assume that "political obligation is voluntarily undertaken and is grounded on this voluntary acceptance itself . . . " (Raphael 1970, p. 102). Obligation depends on something the person does, either some physical or mental action. But the two final theories he reviews, justice and general interest or common good, "concentrate simply on the purposes of the state . . . " (*ibid.*). On the two final theories, a person is morally obligated, if he is, as a consequence of some objective facts about the function of the state and how well the state he is in fulfills those functions. Nothing about the life of the agent-action, will, choice, preference, desire, agreement, activity, and so on-necessarily forms part of the story of how he comes to have this obligation, as it does do on the first three types of theories. Let's call the first three theories "subjective"; the last two, "objective".

Supervenience is an objective relation in much the same sense. Given the way the individual mental facts are (including facts about agents' non-social actions), the social facts are just the way they are. In particular, note that social constructivism gets no toehold on this picture. Agents collectively don't construct the social world, if "construct" is to retain any of its sense that relates it to action or activity. Unlike a supervenience view, constructivism is subjective in the sense that what the social world is like depends not only on facts about the mental lives of agents but also on something those agents do.

Understood in this way, supervenience certainly has its limitations. It has become clear since supervenience was first introduced into the contemporary philosophical

literature that it is a very weak relationship, however much it might be modally strengthened. It is basically a co-variance relation; it simply asserts that the social varies with the individual (and physical) and not vice versa.[3] (Or that the mental varies with the physical and not vice versa.) Supervenience fixes one set of facts as the independent element and claims that a second set varies with that first set.

Some refer to supervenience as a dependency relation, and that is so: the mental (or physical) is the independent variable; the social (or the mental) is the dependent variable. But the idea of dependency when used in this context should not be taken to imply more than that. Given the sun's height, if one makes the length of the flag pole's shadow the independent variable, the height of the flagpole will be the dependent variable, in the sense that the latter's value will depend on the values of the former two. But there is no causal relation running from the length of the shadow to the height of the pole nor does the former explain in any sense the latter.

One form of supervenience, the strongest form, is identity: if social facts were identical to individual mental facts or sets thereof, then the social facts would supervene on the individual mental facts and the identity of the two sets of facts certainly explains why they co-vary, i.e., because they are the same fact(s). In the case of identity, there is nothing more to explain than why we have two ways of talking about the same things. We will need to tell a story about why we have two discourses rather than one, what purpose one fulfills that the other does not. More on this later. (Of course, in the case of identity, we need a separate condition to rule out symmetric supervenience.) But we have already assumed that reductive identity of the social to the mental will fail.

Some relations are not just dependency relations but what we might call determinative: one fact or set of facts explains another because the first makes the second occur, happen, or whatever. Parts make a whole what it is. Causes bring about their effects. Determinative here does not mean deterministic. Even if causation is probabilistic, when c probabilistically causes e, c determines e, brings e about or makes e what it is. Kim's early paper in this area ran together the ideas of dependency and determination (Kim 1974). In that paper, Kim wisely never included supervenience as a determinative relation. In terms of the distinction between dependency and determinative relations, I would classify supervenience as a dependency relation but not as a determinative relation.[4] The subvenient base facts do not *make* the supervenient facts what they are in the way in which causes make their effects what they are or parts make the whole of which they are the parts what it is.

The point below about explanation really follows from the above remarks. When the two sets of facts, the subvenient mental and the supervenient social, are not identical, because supervenience is not a determinative relation, there is nothing in the mere fact of supervenience that obviously explains *why* two sets of distinct

[3] Again, I exclude identity from this claim, although it is the strongest form of supervenience.

[4] I once argued the opposite view, in *Explaining Explanation*, 2012 (second edition). I hereby recant.

facts co-vary, *why* the one is the independent variable and the other the dependent variable. Many writers have noted that what supervenience so understood omits is any explanatory account of why two sets of facts co-vary. On Currie's and List and Pettit's view, it may be true that the social varies with the mental, and not vice versa, but if this is all that can be said, it remains a mystery why this should be so.

Many writers have pointed out how supervenience between two distinct sets of facts by itself leaves an undischarged obligation to resolve a mystery. Two such are Kim and Horgan but there are many others (see, for example, Block 1997). Kim pointed this out in "The Mind-Body Problem: Taking Stock After Forty Years": "... the mere fact ... of mind-body supervenience leaves open the question of what grounds or accounts for it-that is, *why* the supervenience relation obtains ..." (Kim 1997, pp. 189–190). Horgan's goal is to strengthen the supervenience relationship so that there is not only co-variance between supervenient and subvenient facts, but the fact that there is this supervenience (what he calls a "supervenience fact") should be explicable in some way consistent with the subvenient base level (Horgan 1993). This is what he calls "superdupervenience". There are three things to consider on this view: (1) the subvenient base facts (in our case, the individual mental facts and the physical facts); (2) the supervenient facts (the social or institutional facts); and (3) the fact that (2) supervenes on (1). Horgan's claim, for example, for physicalism, is that some account of (3) must be forthcoming that is acceptable to a physicalist. *Pari passu*, a Horgan-like position on our issue is that some account of the fact that the social supervenes on the individual must itself be acceptable to the individualist.

So Horgan's proposal would say of a view like Currie's or List and Pettit's that the view must be strengthened to include a non-social or non-institutional account or explanation of why the social or institutional supervenes on the individualistic base or subvenient level. Horgan himself is dubious that this can be done for the case of the mental, that there can be a materialistically acceptable account of why mental facts supervene on physical facts and he is tempted, as a result, by irrealism about the mental as a way out of the dilemma. What he means by that is this. The difficulty in meeting the requirement for an explanation of the supervenient by the subvenient arises when one presupposes that these are two sets of ("objective") facts. But if the supervenient lacks reality, the explanation is more straightforward. Horgan's example here is R. M. Hare's account of the supervenience of the moral on the non-moral. In its bare bones, Hare is a non-cognitivist. Hare does not assume that there is a set of moral facts distinct from the non-moral facts. Hare talks of the purposes to which moral discourse is put, and it is moral discourse ("the language of morals") that supervenes on non-moral discourse. Hare's view of morality is an irrealist (or non-cognitivist) view and the supervenience thesis is one about two discourses, one factual and the other used for teaching standards, not two areas or realms of reality. The irrealist view comes into its own in the explanation of the purposes of the second discourse, the supervening discourse, a discourse that at bottom is about the same things the first discourse is about.

Does social constructivism provide any explanation of why the social supervenes on the mental or individual and if so, is it a realist or irrealist account? To answer this general question, I want to look briefly at John Searle's *The Construction of*

Social Reality (Searle 1995; subsequent page references in the text are to his 1995, Searle 2010 does not contain any changes in his views relevant to the discussion of this paper). (Note that I am focusing on his account of institutional facts, not social facts, of which the former is merely a proper subset.) I should add that I am not especially interested in a detailed exposition or critique of Searle's discussion. I am hoping to learn some general lessons about the relationship between the social and the individual/mental. Searle's own position is complex and highly nuanced. I will ignore much of that complexity.

When he begins his discussion, Searle speaks as if he is proposing a supervenience account: at the level of types, he says, thinking makes it so: "But where the type of thing is concerned, the belief that the type is a type of money is constitutive of its being money . . . " (1995, p. 33). "If everyone always thinks that this sort of thing is money, . . . then it is money . . . And what goes for money goes for elections, private property, wars, voting, promises, marriages, buying and selling, political offices, and so on" (Searle 1995, p. 32). Searle is usually careful to say that the presence of these beliefs is a necessary condition for the obtaining of a type level social fact, although his use of "constitutive" suggests a stronger position, that the presence of the beliefs is both necessary and sufficient. One might be forgiven for thinking that what Searle is saying is that something's being money supervenes on everyone's believing that it is money. "Constitutive" is this context would just be a metaphor for supervenience.

But constitution as supervenience is certainly not Searle's considered view. There is a second and far more dominant thought in Searle and it has run through all of his writing on this topic. As I read Searle, it is the idea of a constitutive rule that finally converts, as it were, individual or brute (as he says) facts into institutional ones. He says that constitutive rules have the form: X counts as Y in C. Searle does not elucidate the idea of "counting" or distinguish it clearly from identity. It may be that "X counts as Y" and "X = Y" say the same thing. Might "this piece of paper counts as money" be written alternatively as: "this piece of paper is money"? I think not. "Counting as" has a subjective ring to it that "is identical to" lacks.

It is important not to reify constitutive rules, or indeed rules of any kind. Constitutive rules have their reality rooted in the mental life of agents as well: constitutive rules themselves supervene on individuals' beliefs: for example, the existence of the constitutive rule that "X counts as Y in C" supervenes on social agents believing that X is or counts as Y in C. Statements about the beliefs, desires and actions of individual agents give the truth conditions for the existence of a constitutive rule.

In terms of Horgan's discussion, does this show that there is an explanation of the fact that the social supervenes on the non-social in a way that is acceptable to the individualist? The explanandum is: why do these supervening facts supervene on these subvenient facts? The beliefs held by agents that X is or counts as Y in C, and the non-social actions of those agents, are simply further beliefs and actions to be included within the subvenient set of facts, since they are simply further facts about the beliefs and actions of individuals. One cannot explain *why* the social supervenes on individual mental facts by invoking facts that are themselves part of the supervenient base facts. The explaining facts need to be acceptable to the

individualist but they cannot be part of the explanans itself. To include them both in the subvenient base and to invoke them as explanatory of why the supervenient supervenes on the subvenient would be a form of self-explanation. So I cannot see how invoking constitutive rules by itself could allow us to discharge the explanatory burden that Horgan and Kim, among others, require.

But constitutive rules are not where the story stops. After all, all rules have "purchase", as it were, on some people and not others. Yankees did not recognise Confederate money; Sharia Law does not bind Hindus. Searle speaks of rules as arising from agreement or acceptance or consent and I take this to mean that their purchase on some individuals but not others arises from the agreement or consent of the former but not the latter. "The Y term has to assign a new status that the object does not already have just in virtue of satisfying the X term; and there has to be collective agreement, or at least acceptance, both in the imposition of that status on the stuff referred to by the X term and about the function that goes with that status" (Searle 1995, p. 44). Although pp. 47–48 certainly lower the barrier for something to count as agreement in Searle's sense, it has been a theme in his writing from its first presentation (Searle 1964) that some subjective element enters here, variously described: agreement, acceptance, consent, commitment, undertaking. The movement from individual to institutional requires some sort of activity, in however an attenuated sense of that term, on the part of the social agents involved.

It is virtue of this that Searle's account can be called "constructivist". This at least can give some content to the rather vague idea of construction. Social agents construct the social world via their acceptance of or consent to these constitutive rules. I think this is a clear meaning we can give in general to the idea of construction: construction can be cashed out as a kind of agreement or consent, etc. Perhaps we can then understand the agreement and consent story as the super-duper explanation for the supervenience linkage. It is agreement or consent or whatever that explains why this piece of paper counts as money, why its being money supervenes on its being paper plus the agents' beliefs that it is money. More fully: it is the agreement and consent that this piece of paper counts as money that explains why this piece of paper does count as money. (The reappearance of "money" does not itself entail any circularity: see Searle, pp. 52–53.)

But how does this finally meet the requirements of superdupervenience? Recall that the explanatory facts must on the one hand be acceptable to the subvenience point of view but on the other cannot themselves be part of the subvenient base facts. The explanatory facts are meant to be facts about the consent or agreement by the agents to things counting as something. The idea of consent or agreement to constitutive rules raises the same problem as did the beliefs, desires and actions that ground constitutive rules. If consent or agreement is a social action, then consent or agreement won't be acceptable to the individualist as an explanation. If consent or agreement is a non-social action or a mental state, then it belongs in the subvenient base and as such can hardly explain why the supervenient supervenes on the subvenient.

There is an interesting objection to Horgan's way of thinking about superdupervenience. The objection is meant to show that the demand for explanation is

impossible to fulfill in any event. As Lynch and Glasgow argue, the facts about the supervenience relation ('S-facts), that are meant to explain the supervenience, must either themselves supervene on the subvenient facts or on something else (Lynch and Glasgow 2003, pp. 208–209). It's pretty clear that neither option is attractive. (1) If the S-facts supervene on the original subvenient facts, then we have a new set of supervenient facts (that the S-facts supervene on the subvenient facts), call them S*. We then will face a regress since we will need to explain the new S*-facts. (2) If the S-facts do not supervene on the subvenient facts, then they must be *sui generis*, neither part of the subvenient or supervenient facts and hence pretty mysterious. If the objection is well founded, and if the explanatory demand for supervenience is impossible to be met, it is not clear to me exactly where this leaves the idea of supervenience, but generally it does not leave us much confidence in the idea. We would be left with a co-variance relation between two sets of facts, and in principle with no explanation for that co-variance.

How does this connect with Searle or constructivism in general? The supervenience fact (S) is the fact that this counts as money supervenes on the fact that people accept or consent that this piece of paper is money. But why should this fact be so? What now explains (S) or on what does (S) supervene? After all, this schema is not generally true: the fact that everyone agrees or consents to the fact that p does not generally make p true. Jupiter counts as a planet, and an asteroid does not, and a cow counts as a ruminant and a pig does not, but (once meanings of words are fixed), people's accepting or consenting to its counting or not has no role to play in its being a planet or ruminant or not. Given the meaning of "ruminant", a pig isn't and a cow is, whatever people might accept or consent to.

So what we still lack is any account of why accepting and consenting to something counting as something else in the social case and not in other cases should bring it about that something really does count as or is something else. Just as Horgan was tempted to irrealism given the inability to give a superdupervenient account of why the mental supervened on the physical, I would say that any attempt, along these lines, to explain why something's counting as something else supervenes on peoples' consent and agreement that it should so count will lead to a form of social irrealism. "Counting as" is a more subjective relation than "is identical with", and we can now see why. And that is certainly consistent with the overall import Searle attributes to his constructivist account. So perhaps no great news here.

However, Searle's account inherits many of the same problems of consent theories of political and legal obligation. First, we still need an account of whether these are social or non-social actions and, if the latter, how they can play any role in constructing the institutional or social from the brute. Second, if we speak in terms of consent or agreement, they turn out to be empirically false; at best, there is hypothetical consent or tacit consent: if they had been asked, they would have agreed or consented. Or by doing something (according to Locke, like travelling on the King's highways), it is just as if they had consented.

Most people who count bits of paper as money have never agreed or consented to this-they just do it. So it is false that they have ever consented or agreed. What

we will need is an acceptance (or tacit consent) account, since that is the most that most people do when they treat bits of paper as money. In the normal sense of "acceptance", a person accepts that X is Y only because it is already the case that X is Y. That won't do here, for obvious reasons. So I conclude on this critical note: we need some positive account in terms of acceptance or tacit consent of rules such as "X counts as Y in C", in order to finally judge whether Searle, or any constructivist following a similar pattern of argument, succeeds in adding anything to a standard account in terms of supervenience.

References

Block, N. 1997. Anti-reductionism slaps back. In *Philosophical perspectives, 11*, ed. J. Tomberlin, 107–132. Oxford: Blackwell.

Currie, G. 1984. Individualism and global supervenience. *The British Journal for the Philosophy of Science* 35: 345–358.

Horgan, T. 1993. From supervenience to superdupervenience: Meeting the demands of a material world. *Mind* 102: 555–586.

Kim, J. 1974. Non causal connections. *Noûs* 8: 41–52.

Kim, J. 1997. The mind-body problem: Taking stock after forty years. In *Philosophical perspectives*, ed. J. Tomberlin, 185–207. Oxford: Blackwell.

List, C., and P. Pettit. 2011. The structure of group agents. In *Group agency*, ed. C. List and P. Pettit, 59–78. Oxford: Oxford University Press.

Lynch, M., and J. Glasgow. 2003. The impossibility of superdupervenience. *Philosophical Studies* 113: 201–221.

Raphael, D.D. 1970. Democracy. In *Problems of political philosophy*, ed. D.D. Raphael, 83–112. London/Basingstoke: Macmillan.

Ruben, D.-H. 1984. *The metaphysics of the social world*. London: Routledge and Kegan Paul.

Ruben, D.-H. 2012. *Explaining explanation*, 2nd ed. Boulder: Paradigm Publishers. Chap. 7.

Searle, J. 1964. How to derive an 'ought' from 'is'. *Philosophical Review* 73: 43–58.

Searle, J. 1995. *The construction of social reality*. New York City: The Free Press.

Searle, J. 2010. *Making the social world*. Oxford: Oxford University Press.

The Social Re-Construction of Agency

Katarzyna Paprzycka

1 Introduction

One of the deep roots of opposition to social constructionism is the belief that the very idea of a *social* construction of physical concepts is highly suspect. In this paper, I want to call attention to the fact that such "constructions" can occur in the other direction, as it were, as well. There are concepts that really are social and yet we have a tendency to construe them as mental or psychological concepts.[1] In this connection, I want to call attention to a responsibilist theory of agency. According to responsibilism (of which H.L.A. Hart's ascriptivism is an example), actions do not exist as events in the world. Rather, we should understand the attributions of actions in terms of our holding one another responsible for certain events in the world. One of the advantages of responsibilism is that it can capture rather easily a wide class of forms of agency such as omissions (including unintentional omissions), spontaneous, arational, habitual, automatic actions, as well as slips and mistakes. By contrast, the intentionalist approach to agency, according to which an action is paradigmatically a bodily movement caused in the right way by (or otherwise appropriately related, e.g. by means of teleological relations, to) an appropriate rationalizing mental state (belief-desire, intention, intention-in-action, volition), has to either discount such performances as nonagentive or else has to stretch mental

[1] In calling attention to such mental constructions of what is at roots a social phenomenon, I am not siding with social constructionism. In fact, the debate between social constructionism and realism frequently is really a debate between conceptualism (or antirealism) and realism (see also Hacking 2000), where the modifier "social" does not play a great role.

K. Paprzycka (✉)
Institute of Philosophy, University of Warsaw, Krakowskie Przedmiescie 3,
00-927 Warsaw, Poland
e-mail: kpaprzycka@uw.edu.pl

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_22,

concepts so as to encompass the performances. It is this concept stretching that underlies the "mental construction" in question. From the responsibilist point of view, intentionalism has a tendency to (mis) construe as primarily mental concepts that are, at roots, social.

In Sect. 1, I distinguish various senses of the notion of action and try to make clear the concept I am after. In Sect. 2 I sketch the distinction between the two types of theories of action and show them to have a common root in Aristotle's remarks on what is voluntary. In Sect. 3 I sketch Hart's version of responsibilism and briefly answer some immediate objections (Sect. 4). In Sect. 5 I sketch one way of understanding the notion of control that I take to underlie responsibilism. In Sect. 6 I point out some advantages of responsibilist theories and show how, from the point of view of responsibilism, intentionalism undertakes a "mental construction" of what are ultimately social phenomena.

2 Action as a Unit of Conduct

Perhaps the most fundamental difficulty in analyzing the concept of action is the fact that it plays a significant role in a number of disciplines as diverse as physics, biology, psychology and sociology. As a result, the concept has coalesced a great variety of intuitions. It is thus important at least to try to distinguish some ways of understanding the term "action".

1. There is a concept of *inanimate* action. When a billiard ball thrusts into another billiard ball it acts on the other. To its action, by Newton's third law, there corresponds an appropriate reaction of the other ball. The application of teleological concepts to inanimate action appears to be only derivative. For example, we can speak of the purpose or function of a piece of a thermostat but its purposefulness is derived from its being designed by someone.
2. We can speak of the actions of various parts of animal bodies. This is the first stage at which non-derivative teleological concepts find application. The liver's excreting bile, the heart's pumping are examples of what one might call *purposeful* movements or actions.
3. The third level is that of *purposive movement* or action. The subjects of our attributions of purposive movements are no longer parts of bodies but rather agentive systems. The movements a spider produces in spinning a web constitute purposive movements. In this sense also, a drug addict's compulsively taking a shot is purposive (Frankfurt 1988, pp. 76–77). Arguably, sleep-walking, some actions performed under hypnosis, as well as the little movements one performs to alleviate muscle pain in one's sleep are purposive. So are feeding the cat, conversing, looking out of the window, walking through a forest.

4. The latter but not the former examples belong to a more restrictive category of *intentional movements*. A movement is intentional just in case there is some description under which it is intentional. The category of intentional movements is an extensional category – it picks out a class of events. As such, it is a very different concept from the concept of intentional action, which is an intensional concept (Anscombe 1957; Davidson 1971). Both intentional and unintentional actions, as they are usually understood, are intentional movements in this sense.[2]

It is not uncontroversial to sharply distinguish the category of intentional movements from the category of purposive movements. One might treat the distinction to be one of degree rather than principle. Yet many of the examples relevant here are at least very different from the examples of purposive movements. So when one goes to a rally, one performs an action of a different sort than if one went there in one's sleep.

5. The fifth sense of "action" derives from the idea of an agent's overall conduct. Someone's conduct includes her intentional and unintentional doings but also intentional and unintentional not-doings (omissions). When we inquire after a person's conduct during a rally, say, we will be interested in the things the person said and did as well as the things that he omitted to say or do.

The concept of action as part of an agent's conduct has not been at the forefront of philosophers' concern with agency.[3] Most of the debate has centered around the concept of action in the sense of purposive and/or intentional movement. This is, among other things, because intelligence and reason are most clearly manifested in our acting intentionally. But the philosophical focus on "intelligent agency" should not lead one to think that there is nothing interesting about action but for its rational significance. In fact, there are psychological categories that pertain to our conduct rather than merely to our intentional behavior. The most important among them is the concept of character. Character comprises not only our agentive voice – active intentional rational excursions into the world – but also our idleness, passivity, thoughtlessness, carelessness, forgetfulness – our agentive silence, as it were. The responsibilist is best viewed as trying to capture the fifth sense of the concept of action.

[2]This usage of the term "intentional movement" is not widespread. The term is introduced by Frankfurt 1988). In view of the dispute between minimalism (e.g. Davidson 1971) and moderationism (e.g. Thomson 1977) in the ontology of action, the claim would have to readjusted. Moderationists would insist that intentional movements are parts (possibly proper parts) of complexes with which intentional (unintentional) actions are identical.

[3]The most obvious exception is H.L.A. Hart (1951) who frequently speaks of the "philosophy of conduct," intending to cover both actions and omissions (including unintentional ones) by the term.

3 Two Types of Action Theories

Two general types of theories of action could be seen to have a source in Aristotle's remarks on the voluntary. Aristotle says:

> What comes about by force or because of ignorance seems to be involuntary. What is forced has an external origin, the sort of origin in which the agent or victim contributes nothing – if, e.g. a wind or human beings who control him were to carry him off.[4]

These remarks are suggestive of a certain picture of what it means for a performance to be a mere happening rather than an action:

(N_e) The agent's φing was a mere happening (nonaction) iff "external forces" caused him to φ.

and a corresponding picture of what it means for a performance to be an action:

(A_i) The agent's φing was an action iff "internal forces" caused him to φ.

(N_e) can be thought to capture the central thought of responsibilism while (A_i) captures the central thought of intentionalism. The ideas of "internal forces" and "external forces" are treated as a stand-in for a more detailed account.[5]

Despite appearances, however, (N_e) and (A_i) are not so easily reconcilable with one another. In particular, it is worth stressing that (A_i) does *not* follow from (N_e), nor (N_e) from (A_i). In order to establish logical relations between them, one would have to relate the concepts of mere happening and action, on the one hand, and of "external" and "internal forces," on the other.

The former move is the least problematic of the two: we can view the concepts of action and nonaction as complementary (relative to a class of performances[6]):

(A-N) A performance is an action just in case it is not a mere happening (nonaction).

Given (A-N), we can establish that what follows from (N_e) is:

(A_e) The agent's φing was an action iff it was not caused by "external forces."

Analogically, what follows from (A_i) is:

(N_i) The agent's φing was a mere happening iff it was not caused by "internal forces."

---

[4] Aristotle, *Nicomachean Ethics*, trans. Terence Irwin (Indianapolis: Hackett 1985), 1110a1–4.

[5] One must remember to avoid simple-minded interpretations here. The distinction is not (as suggested by the form of words Aristotle sometimes uses) between forces outside and inside the agent, for there can be the wrong kind of forces inside the agent (spasms, e.g.), and there may be the right kind of "external" forces (e.g. when someone helps an old person through the street). See also (Frankfurt 1988).

[6] The use of the term "performance" is technical. It is constructed in such a way as to encompass both actions and nonactions – the winkings and the blinkings, the arm raisings and the arm risings, falling off the stairs and running down the stairs etc.

The question is whether (A_i) and (N_i) (as well as (N_e) and (A_e)) can be reconciled. They can provide that the following claim is true:

(i-e) "internal forces" caused α to φ iff "external forces" did not cause α to φ.

The status of claim (i-e) has not been given enough attention. What is to preclude the possibility that both types of causes occur at the same time (see Paprzycka 2013)? Traditionally, philosophers of action have adopted two strategies. One might begin with elucidating the idea of what it means for the "internal forces" to cause an agent's performance (A_i) and then explain what it means for the agent's performance to be a mere happening in terms of (N_i). This strategy is typical of intentionalism. What this special origin is will depend on the theory in question. The causal theory of action could be seen as paradigmatic for this approach, and one which has reigned over philosophers' intuitions about action for a very long time. On that view, an action is an event that is caused in the right way by appropriate mental states (e.g. Brand 1984; Davidson 1963, 1973; Mele 1992, 2003; Searle 1983). But even many of the challengers to the causal theory share the basic intuition. On the ever more popular agent-causal views (e.g. Chisholm 1976; Lowe 2008; O'Connor 2000), an action is an event that also has a special origin – it must be caused by an agent. Volitionalist approaches (e.g. Ginet 1990) likewise follow this general line of thinking.

Alternatively, one might begin with the idea of what it means for "external forces" to cause an agent's performance (N_e) and then explain what it means for the agent to act by appealing to the absence of such forces (A_e). This is the strategy pursued by responsibilism. From its point of view, the idea of "internal forces" causing the performances is a hypostatization of the absence of such causation by "external forces." In other words, responsibilists take (i-e) to be providing a reductive definition of what it means for the "internal forces" to be in operation. The proper theoretical work is done by the notion of the absence of "external forces." This kind of approach to action has been proposed by H.L.A. Hart (1951), and unfortunately much forgotten since (though see Paprzycka 1997, 2008; Sneddon 2006; Stoecker 2001, 2007). On Hart's view, the notion of action is best understood as a complement to the notion of nonaction (of a mere happening). Nonactions in turn are understood in terms of the presence of defeating conditions (e.g. spasms, ticks, etc.).

What is characteristic of responsibilism is that we are entitled to attribute actions to one another by default, as it were, but we are committed to withdrawing the attribution if certain defeating conditions are present. On the intentionalist approach, by contrast, we are entitled to attribute actions to one another only if we have reasons to believe that the event has had the appropriate origin.

4 H.L.A. Hart's Theory of Action

Hart argued that the concept of action, like the concept of property, is essentially normative and social in that it presupposes accepted rules of conduct. Just as it does not make sense to think that a statement like "Smith owns this piece of land" is a sentence that is "concerned wholly with an individual" (Hart 1951, p. 161), so statements like "Smith did it" likewise should not be understood as describing one individual only. The main purpose of action statements is to attribute ("ascribe" in Hart's language) responsibility for certain events[7] to individuals on the basis of generally accepted rules of conduct.

Hart proposes that claims like "John broke the glass" not be interpreted as describing an action but rather as ascribing responsibility to the agent (here: for the glass breaking). Action claims are ascriptive rather than descriptive. They are never true or false; they may only be appropriate or inappropriate in view of relevant conditions (Geach 1972; Paprzycka 1997; Stoecker 2007). Transposed from the formal into the material mode, there are no actions among the ontological furniture of the world.

What distinguishes actions from mere happenings, on Hart's view, is not any ontological fact but rather the appropriateness of ascribing responsibility for events in certain conditions (when we intuitively think of them as actions) and the inappropriateness of ascribing responsibility for events in other conditions (when we intuitively think of them as mere happenings). This is what it means to say that the distinction between actions and mere happenings is normative in nature. But this is not yet to give an account of the distinction. In fact, Hart never provides such an account but rather notes that there are conditions that determine whether it is appropriate or inappropriate to ascribe responsibility to the agent.

The structure of action attribution is characteristically defeasible. First, there are, in Hart's terminology, positive conditions that establish the prima facie applicability of the responsibility attribution. In our example, such conditions include John's arm moving in such a way as to break the glass. Second, there are negative (defeating) conditions that defeat the prima facie appropriateness of ascribing responsibility to the agent. Such conditions include John's arm moving because of a spasm, say, or John's arm moving as a result of someone, Mary say, taking it and guiding through the motion. When we learn that it was Mary who took John's arm and smashed the glass with it, we would no longer attribute responsibility to John for breaking the glass. Mary's taking John's arm counts as a defeating condition.

It should be pointed out, however, that defeating conditions can themselves be defeated. Suppose that John held Mary at a gunpoint and ordered her to take his arm

[7]It is not exactly clear on Hart's account what ontological category the variable x ranges over in the expression "α is responsible for x". Some critics (e.g. Pitcher 1960) have charged Hart with the view that the variable ranges over actions, thus rendering Hart's account circular. But, most of the time (except for a noncommittal statement on the first page of his paper), Hart is quite careful not to talk this way. The variable could be interpreted as ranging over events or events under a description. (For details, see Paprzycka 1997, ch. 2. See also Sneddon 2006.)

and break the glass with his arm. Such a condition defeats the original defeating condition. As a result, it is reasonable to hold John to be responsible for the glass breaking after all. He did it in an unusual way but he did break the glass.

This structure allows us to understand the difference between actions and mere happenings or between it being appropriate and it being inappropriate to ascribe responsibility to an agent. It will be inappropriate to ascribe responsibility to the agent if either no positive conditions are present or while the positive conditions are present some (undefeated) defeating condition occurs. It will be appropriate to ascribe responsibility to the agent if the positive conditions occur and no (undefeated) defeating conditions are present.

5 Some Objections

Many objections could be raised to such an approach. Let me consider three.

First, what is one to do with *mundane actions*. Even if John may be responsible for breaking a vase by raising his arm, what if nothing of any (moral or legal) significance happened as a result of John's movements. He just raised his arm. Likewise, he just sang in the shower. Surely, we want to say that these are John's actions but it seems a stretch to think that he is responsible for them.

A responsibilist can think that John is neither legally nor morally responsible for singing in the shower. There are a couple of options open here. One would be to say that the concept of action is built over the concept of responsibility but applies more widely than does the former. We speak of actions not just when we hold agents responsible but when we could hold them responsible, i.e. when it would be appropriate to hold them responsible in appropriately changed circumstances. It does not matter that nothing of substance hangs on John's singing in the shower as long as *if it did*, it would be appropriate to hold John responsible. Another sort of response would be to propose a concept of the right sort of responsibility, which identifies those elements in the concept of responsibility that are relevant to the attribution of action. This is the path I have followed in (Paprzycka 1997) by proposing the concept of practical responsibility (briefly sketched in Sect. 6).

Second, actions are usually understood as performances that are intentional under some description (Anscombe-Davidson thesis), yet the idea of *intentional action* seems completely lost in responsibilism. It is true that the idea of intentional action loses its center-stage role. There is thus no conceptual pressure for the responsibilist to uphold Anscombe-Davidson's claim. However, this is actually an advantage of the stance rather than its disadvantage.[8] One of the notorious features of intentionalist views is that because they are under pressure to uphold Anscombe-Davidson's thesis, they are also under pressure to stretch the concept of intentional action in ways that we might otherwise resist.

[8]It should be noted that there are ways of developing responsibilism where the notion of intentional action does play quite a significant role (see e.g. R. Stoecker 2001, 2007).

On the other hand, although the concept of intentional action does not ground the concept of action on responsibilist views, there remains a need to understand what intentional actions are. Conceivably, a responsibilist could even adopt a causalist account of intentional action (e.g. Mele and Moser 1994). The main thrust of responsibilism is to deny the claim that the notion of intentional action is conceptually prior to the notion of action. The responsibilist claims that the reverse is true.

Third, there seems to be *no role for beliefs, desires* and other mental states. Again, the responsibilist will say that the expectation that mental states play any role in an account of action is to be rejected with intentionalism. This is not to say that she might not return to them in an account of action explanation. Arguably, however, she is not forced to limit herself to the agent's mental states in an account of action explanation. In fact, the responsibilist may adopt a non-individualist account (see e.g. Baier 1985; Collins 1987; Nowak 1987, 1991; von Wright 1983; Paprzycka 1998, 2002; Schmid 2008, 2009; Wilson 1989), according to which one may very well appeal, for example, to others' mental states in explaining the agent's action (e.g. Susie went to the store because her mother wanted some carrots, Joe shocked the victim because he was told to do so by the experimenter, etc.).

Last but not least, there is what deserves to be called the fundamental problem. Usually we think that we are responsible for what we do. Usually we take this to mean that the notion of action is conceptually prior to the notion of responsibility. We need to settle what was done, first, in order to determine whether we are responsible for it. Yet the responsibilist claims that this logical order should be reversed and proposes to build a notion of action on the basis of the concept of responsibility. It looks like the enterprise is bound to be circular (see e.g. Pitcher 1960). Contrary to appearances this is not a devastating objection. The concept of responsibility is not homogeneous. Particularly helpful in this regard is K. Baier's (1980, 1987) distinction between backward-looking concepts of responsibility (answerability, culpability, liability), which presuppose that the agent has something (possibly an action) to answer for, and forward-looking concepts of responsibility (task-responsibility), which do not presuppose that the agent has done something. In fact, the account of practical responsibility sketched below involves the latter notion. It is beyond doubt, however, that the objection has to inform the responsibilist theorizing and keep him on his toes.

6 Practical Responsibility

The general structure of responsibilist accounts can be perhaps cast in the following way. Our attributions of actions of φing to agents presuppose that we take agents to be practically responsible for φing. Agents are, in turn, practically responsible for φing when they are in control of φing or when φing is within their power. When α is practically responsible for φing and φs, α's φing is an action.

Normally we take ourselves to be – and indeed are – in control of various ordinary activities (raising arms, moving feet, reaching for glasses, pouring milk, running,

writing, etc.). Arguably, we differ in the repertoires of activities we are in control of – some of us can juggle, dance or raise arms in graceful ways, others cannot sing, cannot write or may be incapable of walking. However, the very application of the concept of action in our social lives presupposes that the range of activities that we are in control of overlaps quite a bit. The understanding of the idea of what it is to be in control of φing includes at the very least that the agent is capable of reliably fulfilling the task to φ (and the task not to φ).[9]

Defeating conditions – or Hart's negative conditions – are conditions that affect the agent's control of an activity. When someone breaks a leg, she is no longer in control of running or even walking briskly. When someone acquires a tick, he cannot be relied upon to wink as a sign to start a revolt.

The heart of the responsibilist account involves understanding these key concepts, what it is for the agent to be in control of φing and how various conditions can affect this state. Once the responsibilist demonstrates that most agents indeed are in control of φing, where "φ" ranges over "common" action types, he entitles himself to the characteristic feature of responsibilist theories, viz. that such action types lie within the agentive purview by default. With respect to them, we are *guilty of being agents until proven innocent*, so to speak.[10] According to the responsibilist, one does not need to demonstrate any special origin for a performance to count as an action as long as it is a performance of a type that falls within the range of those that are within the agent's power or control. Of course, such action attributions are defeasible – they are sensitive to the occurrence of any circumstances that affect the agent's control of those types of actions.

While it is beyond the scope of this paper to go into the details of responsibilist positions, I want to briefly provide a proposal for an elucidation of the relevant idea of "what it is to be within the agent's power or control" in terms of the idea of reasonable normative expectations (Paprzycka 1997, 1999). I want to stress, however, that this proposal is not inherently tied to responsibilism. It is one way of developing the position but probably not the only way of doing so.

Two intuitions seem to be central to our idea of what it is to be within the agent's power to do something. First (the success intuition), the agent must be able to succeed in φing. It is not within the power of a 1-year old to write a novel, or of someone who broke a leg to run a marathon. Second (the difference intuition), the agent must be able to make a difference – it is not within our power to make the sun shine, or to hold breath for an hour.

One way to capture these intuitions is to envisage an ideal test to which an agent could be subjected. We give the agent a series of tasks, to which he responds in the best possible way; we are assuming, in other words, that he is cooperative, that

[9]In a remarkable paper, Baier (1970) has sketched an approach to action where the idea of a task plays a central role. The account sketched owes a lot to hers.

[10]The thought that some performances acquire the status of actions (indeed intentional actions) by default is present in Brandom's account (see esp. Brandom 1994, pp. 257ff). While Brandom officially adheres to a causal account of action, this thought makes him an implicit responsibilist.

Table 1 Possible result patterns of a simplified test sequence

	Task: φ	Task: not-φ
(i)	pf-fulfilled (φ)	pf-fulfilled (not-φ)
(ii)	pf-fulfilled (φ)	pf-frustrated (φ)
(iii)	pf-frustrated (not-φ)	pf-fulfilled (not-φ)
(iv)	Other	

there are no other designs, intentions, expectations in play, the agent is at ease, under no pressure, etc.[11] The tasks are of two kinds, to φ and not to φ, and they are interspersed randomly in a series.

Three situations are of special interest. Suppose that an agent systematically pf-frustrates[12] the expectation to φ (situation (iii) in Table 1). When he is expected to φ, he does not. In such a case, it would be unreasonable*[13] to expect of the agent that he φ. The agent cannot succeed in meeting the task. Suppose that the agent regularly pf-fulfills the expectation to φ but also regularly pf-frustrates the expectation not to φ (ii). What this will mean is that the agent φs indiscriminately. In such a case, we would tend to think that the agent's φing is not up to him, that the agent cannot make a difference, and hence that it would be unreasonable* to expect of him that he φ. This configuration would obtain if we expected of the agent that he breathe, for example. Finally (i), when the agent pf-fulfills all the expectations (when expected to φ, the agent responds by φing, when expected not to φ, the agent responds by not φing), we would tend to think that φing and not φing are "within the agent's power," that it is reasonable* to expect of the agent that he φ.

We can understand the success and difference intuitions accordingly:

Success condition: It is prima facie unreasonable* to hold α to a (normative) expectation to φ if were α subjected to the test (with all of its conditions satisfied[14]), the expectation to φ would be pf-frustrated.

[11]This is an idealizing assumption. I am making it in order to sharpen the intuitions at stake.

[12]The notions of prima-facie (pf-) fulfillment and pf-frustration are introduced in part to prevent the circularity problem that affects responsibilist accounts. One might, for example, argue that an expectation to raise an arm is not fulfilled when the arm rises of its own accord, by accident, in a spasm, etc. An account of action that used such a concept of expectation fulfillment would be rightly charged with circularity. That is why, at this stage, I appeal to a very liberal notion of pf-fulfillment that includes not only actions but also nonactions as pf-fulfilling an expectation.

[13]I mark the notion of reasonableness with an asterisk to note that it is a theoretical concept, which captures but one dimension of our ordinary concept. In particular, our concept of what it would be reasonable to expect of another person also includes a normative component (it would be unreasonable of me in this sense to expect of my neighbor to mow my lawn even though it would be within his power to do so), which I am ignoring here.

[14]I take the condition "α is subjected to the test" to be synonymous with the expression "α is subjected to the test and α fulfills all the conditions of the test, i.e. is cooperative, at ease, under no pressure, with no other intentions or expectations in play".

***Difference condition**: It is prima facie unreasonable* to hold α to an expectation to φ* if were α subjected to the test, the expectation to φ would be pf-fulfilled but the expectation not to φ would be pf-frustrated.

It is possible for a normative expectation to be prima facie reasonable* (e.g. to run a race) but for certain conditions to occur (e.g. a broken leg) that would defeat its reasonableness*. A person who was subjected to the test in the presence of such conditions would respond quite differently. We can distinguish hindering and compelling conditions accordingly:

Condition C is a **hindering condition** with respect to an expectation to φ iff were α subjected to the test in condition C, the expectation to φ would be pf-frustrated.

Condition C is a **compelling (or forcing) condition** with respect to an expectation to φ iff were α subjected to the test in condition C, the expectation to φ would be pf-fulfilled but the expectation not to φ would be pf-frustrated.

An example of a hindering condition with respect to an expectation to run a race would be breaking a leg. An example of a compelling condition with respect to an expectation to walk would be being forced to do so by another person. We may perhaps distinguish yet a third type of condition, whose occurrence would lead a person to lose control in yet another sense. The person who normally responds reliably to expectations to touch the table and not to touch the table might simply become erratic. Arguably a person with some neural condition or who has taken drugs might fit this kind of pattern. A person in this condition subjected to the test would not exhibit a pattern of responses falling neatly under rows (i)–(iii) but rather fall in row (iv). Such a condition might be called a disabling condition.

An agent α is practically responsible for φing at time *t* if: (a) the expectation to φ is prima facie reasonable*, (b) at time *t*, no (undefeated[15]) defeating condition (with respect to the expectation that α φ) is present.

We can use the notion of practical responsibility in the responsibilist account of action. When John's arm rises it will count as his action as long as he is practically responsible for raising his arm at this moment, or as long as it would have been reasonable* to expect of him that he raise his arm at this moment. If John is like most of us then if we were to subject him to the test (as long as its conditions are satisfied, i.e. John is cooperative etc.), his performances would fit the pattern (i). In other words, the expectation of John to raise his arm is prima facie reasonable. At the same time, as long as no undefeated defeating occurs, the expectation to raise an arm continues to be reasonable*. So as long as John is practically responsible for raising his arm (at *t*) and his arm rises (at *t*), we can attribute to John the action of raising his arm. The attribution does not depend on our knowledge of what mental state John was in at all.

[15]Defeating conditions can be defeated by other conditions and those conditions can be defeated by still other conditions.

7 The Social Reconstruction of Agency: The Case of Intention-in-Action

Central to the intentionalist view of agency is the thought that agents' bodily movements be appropriately related to agents' reasons (often conceived of as mental states that rationalize the actions). Central to the responsibilist view is the thought that the agent has an appropriate sort of control, the lack of which would exclude responsibility. Responsibilists take competent agents to perform actions by default. No special account needs to be given to count an agent's walking movements, for example, as an action. The agent need not have a reason to walk (though she might), a desire to walk (though she might), an intention to walk (though she might) or intention-in-action or volition or … The walking movements of a competent and skilled walker will count as an action of walking whatever its causes and whatever the mental attitude of the walker, as long as the walker is in control, i.e. as long as nothing defeats reasonableness of the expectation of the agent that she walk.[16]

In giving this sort of account of agency, we can give a more accurate account of some types of agency, without having to venture to search for intentional states, the postulation of which often seems just arbitrary. Consider the introduction of the notion of intention-in-action by Searle. He first expresses the belief that there are no actions without intentions (Searle 1983, p. 82). Then he argues that because there are cases of actions which are not preceded by a prior intention, so in such cases the intention must be *in* the action.

In other words, one starts with the theoretical claim that for a performance to be an action, it *must* be suitably related to an intention. But there are actions which are not done on (prior) intentions. In fact, there are spontaneous actions, actions done for no reason (like one's humming a tune under shower). Rather than taking them as a problem for the theory, one saves the theory by postulating that in case of those performances there are intentions after all, they are just concurrent with actions – intentions-in-action. Thus introduced the concept of intention-in-action is an *ad hoc* device introduced to save the theory that actions must be related to intentions. Moreover, we may suspect that the concept of intention-in-action is really parasitic on the concept of action. We have a much firmer grip on the concept of action than we do on the concept of intention-in-action. To the extent that the notion of intention-in-action serves as an *explanation* of what an action is, it is

[16]For similar reasons, the responsibilist is open to the idea of letting in as actions cases that are notoriously problematic for the intentionalist because the requisite internal make up seems lacking: arational actions (Hursthouse 1991), habitual actions (Pollard 2003, 2006), nonintentional actions (Chan 1995), unintentional omissions (Smith 1984, 1990), mistakes and slips (Peabody 2005). At the same time, cases of antecedential wayward causal chains do not present any special problems (see also Paprzycka 1997, 2013).

based on theoretical trickery. We have first coined the notion of intention-in-action on the basis of our intuitions about the concept of action. Then we use the conjured concept to explain our concept of action.

Searle uses the following example to make his suggestion intuitive:

> Suppose you ask me 'When you suddenly hit that man, did you first form the intention to hit him?' My answer might be, 'No, I just hit him.' But even in such a case I hit him intentionally and my action was done with the intention of hitting him. I want to say about such a case that the intention was in the action but that there was no prior intention (Searle 1983, p. 84).

It would indeed be unreasonable to be suspicious of the possibility that an intention may arise on the spur of the moment. But to allow such a possibility is not yet to buy into the claim that an intention in action, understood as a mental state, is present in all cases of action. Consider the following case.

Suppose that I walk down the street, engrossed in thoughts, picking leaves from nearby bushes as I walk by them. You catch me doing this. I may have not even realized that I was doing so. You ask "When you picked those leaves, did you form an intention to do so?" I am very likely to answer just like Searle did – "No. I was just picking them, I suppose." Did I do so *with the intention* of picking the leaves? This is a very strange question to ask in this situation. If what you mean by that is "Was I picking the leaves?", then my answer will be positive. (It is clear, however, that you are not asking this question since you knew I was picking the leaves before I did.) If you are asking about my intention then I would be inclined to answer in the negative. I certainly had *no reason* at all to pick them. I would also oppose the view that I wanted to pick them (Paprzycka 1998, 2002).

Perhaps the intention is just the goal toward which my movements were directed. However, consider a case of an action slip reported by James:

> Very absent-minded persons in going to their bedroom to dress for dinner have been known to take off one garment after another and finally to get into bed, merely because that was the habitual issue of the first few movements when performed at a later hour (James 1890/1983, p. 119).

What are we to say about such a case? The person intends to dress for dinner and thus *intentionally* prepares to go to bed? This is not a case of a change of mind: the person still intends to prepare for dinner and goes on to do so after realizing the mistake. This is not a case of a mistaken belief: the person does not have the belief that going to bed is a good way of dressing for dinner (see also Peabody 2005). Does the person have an intention in action to go to bed? This is what the person is doing. He has no reason to do it. In fact he has reasons not to do it. His movements are directed at going to bed just as they are when he goes to bed with the intention of doing so. The problem is that unlike in the usual case, in this case the agent disowns that goal. If he attends to the situation, he will vehemently deny that this was his intention.

Note too that this is to allow that when one looks at the slip from a psychological or neurocomputational point of view one will be able to find the activation of states that are responsible for certain motor routines (sometimes also referred to as "motor intentions"). But there is a gulf between such states and philosophers' intentions, which rationalize the action and which are responsible for the action's intentional

character. The neurocomputational difference between what happens in the case when someone intentionally dresses for dinner and when someone intends to dress for dinner and slips into preparing for bed may be a matter of degree (perhaps the higher activation of one or two units made the difference). From a philosophical point of view the difference is qualitative – in one case the person acts rationally and intentionally, in the other – the performance is irrational in these circumstances and not intentional.

From a responsibilist point of view, this tension between the notion of intention, which is tied with the rationalization of an action, and the notion of intention, which is responsible for a performance being an action, can be removed. The notion of intention can be reserved for giving an account of the rationalization and intentionality of action. There are good reasons to suppose, however, that not all of our actions are performed for a reason. There are good reasons to suppose that not all of our actions are intentional under some description. All of our actions presuppose, however, that we exhibit the right sort of control, which is prerequisite of our being responsible.

If this account is correct then the intentionalist presents us with a "mental construction" (here: the postulation of intention-in-action) of what is at roots a social phenomenon (here: the absence of defeating conditions). This is indeed one of the complaints of H.L.A. Hart who warned against this kind of construction:

> These positive-looking words 'intention', etc., if put forward as necessary conditions of all action only succeed in posing as this if in fact they are a comprehensive and misleadingly positive-sounding reference to the absence of one or more of the defences, and are thus only understandable when interpreted in the light of the defences, and not vice versa. Again, when we are ascribing an action to a person, the question whether a psychological 'event' occurred does not come up in this suggested positive form at all, but in the form of an inquiry as to whether any of these extenuating defences cover the case (Hart 1951, p. 163).

The responsibilist advocates a view of agency, which is fundamentally social. It is only available within quite complex practices of holding one another responsible. In resisting the idea that agency is a social concept, we have become all too accustomed to thinking that it is a mental concept. As a result, we have found it much easier to postulate all kinds of mental states or processes than trying to find resources already at our disposal when we view it as a social concept. Perhaps it is time to pause and reflect lest we should witness the debate between realism and physical constructionism in the future.

Acknowledgement The work on the paper has been sponsored in part by an NCN grant (DEC-2012/05/B/HS1/02949).

References

Anscombe, G.E.M. 1957. *Intention*. Oxford: Basil Blackwell.
Aristotle. 1985. *Nicomachean ethics*. Trans. Terence Irwin. Indianapolis: Hackett.
Baier, A.C. 1970. The search for basic actions. *American Philosophical Quarterly* 8: 161–170.

Baier, K. 1980. Responsibility and action. In *Action and responsibility*, ed. M. Bradie and M. Brand, 100–116. Bowling Green: Bowling Green University Press.
Baier, A.C. 1985. *Postures of the mind: Essays on mind and morals*. Minneapolis: University of Minnesota Press.
Baier, K. 1987. Moral and legal responsibility. In *Medical innovation and bad outcomes*, ed. M. Siegler, S. Toulmin, F.E. Zimring, and K.F. Schaffner, 101–129. Ann Arbor: Health Administration Press.
Brand, M. 1984. *Intending and action. Toward a naturalized action theory*. Cambridge: MIT Press.
Brandom, R. 1994. *Making it explicit*. Cambridge: Harvard University Press.
Chan, D.K. 1995. Nonintentional actions. *American Philosophical Quarterly* 32: 139–152.
Chisholm, R.M. 1976. *Person and object*. La Salle: Open Court.
Collins, A.W. 1987. *The nature of mental things*. Notre Dame: University of Notre Dame Press.
Davidson, D. 1963. *Actions, reasons, and causes*. Reprinted in Davidson, D. 1980. *Essays on actions and events*, 3–19. Oxford: Clarendon Press.
Davidson, D. 1971. *Agency*. Reprinted in Davidson, D. 1980. *Essays on actions and events*, 43–61. Oxford: Clarendon Press.
Davidson, D. 1973. *Freedom to act*. Reprinted in Davidson, D. 1980. *Essays on actions and events*, 63–81. Oxford: Clarendon Press.
Frankfurt, H. 1988. The problem of action. In *The importance of what we care about*, ed. H.G. Frankfurt, 69–79. Cambridge: Cambridge University Press.
Geach, P.T. 1972. Ascriptivism. In *Logic matters*, ed. P.T. Geach, 250–254. Berkeley: University of California Press.
Ginet, C. 1990. *On action*. Cambridge: Cambridge University Press.
Hacking, I. 2000. *The social construction of what*. Cambridge: Harvard University Press.
Hart, H.L.A. 1951. The ascription of responsibility and rights. In *Essays on logic and language*, ed. A. Flew, 145–166. Oxford: Blackwell.
Hursthouse, R. 1991. Arational actions. *Journal of Philosophy* 88: 57–68.
James, W. 1890/1983. *The principles of psychology*. Cambridge: Harvard University Press.
Lowe, E.J. 2008. *Personal agency. The metaphysics of mind and action*. Oxford: Oxford University Press.
Mele, A.R. 1992. *Springs of action*. Oxford: Oxford University Press.
Mele, A.R. 2003. *Motivation and agency*. Oxford: Oxford University Press.
Mele, A.R., and P.K. Moser. 1994. Intentional action. *Nous* 28: 39–68.
Nowak, L. 1987. Man and people. *Social Theory and Practice* 14: 1–17.
Nowak, L. 1991. *Power and civil society: Toward a dynamic theory of real socialism*. New York: Greenwood Press.
O'Connor, T. 2000. *Persons and causes. The metaphysics of free will*. Oxford: Oxford University Press.
Paprzycka, K. 1997. *The social anatomy of action: Toward a responsibility-based account of agency*. Ph.D. dissertation, University of Pittsburgh.
Paprzycka, K. 1998. Collectivism on the horizon: A challenge to Pettit's critique of collectivism. *Australasian Journal of Philosophy* 76: 165–181.
Paprzycka, K. 1999. Normative expectations, intentions and beliefs. *The Southern Journal of Philosophy* 37(4): 629–652.
Paprzycka, K. 2002. False consciousness of intentional psychology. *Philosophical Psychology* 15: 271–295.
Paprzycka, K. 2008. Sneddon on action and responsibility. *Polish Journal of Philosophy* 2: 69–88.
Paprzycka, K. 2013. Can a spasm cause an action? *Grazer Philosophische Studien* 87: 159–174.
Peabody, K. 2005. Trying slips: Can Davidson and Hornsby account for mistakes and slips? *Philosophia* 35: 173–216.
Pitcher, G. 1960. Hart on action and responsibility. *The Philosophical Review* 69: 226–235.
Pollard, B. 2003. Can virtuous actions be both habitual and rational. *Ethical Theory and Moral Practice* 6: 411–425.
Schmid, H.B. 2008. Plural action. *Philosophy of the Social Sciences* 38: 25–54.

Schmid, H.B. 2009. *Plural action. Essays in philosophy and social science*. Dordrecht: Springer.
Pollard, B. 2006. Explaining actions with habits. *American Philosophical Quarterly* 43: 57–68.
Searle, J.R. 1983. *Intentionality. An essay in the philosophy of mind*. Cambridge: Cambridge University Press.
Smith (Milanich), P.G. 1984. Allowing, refraining, and failing. The structure of omissions. *Philosophical Studies* 45: 57–67.
Smith, P.G. 1990. Contemplating failure: The importance of unconscious omission. *Philosophical Studies* 59: 159–176.
Sneddon, A. 2006. *Action and responsibility*. Dordrecht: Springer Academic Press.
Stoecker, R. 2001. Agents in action. *Grazer Philosophische Studien* 61: 21–42.
Stoecker, R. 2007. Action and responsibility – a second look at ascriptivism. In *Intentionality, deliberation and autonomy: The action-theoretic basis of practical philosophy*, ed. C. Lumer and S. Nannini, 35–46. Aldershot: Ashgate.
Thomson, J.J. 1977. *Acts and other events*. Ithaca: Cornell University Press.
von Wright, G.H. 1983. Explanation and understanding of action, in his G.H. von Wright, *Practical reason*, 53–66. Ithaca: Cornell University Press.
Wilson, G.M. 1989. *The intentionality of human action*. Stanford: Stanford University Press.

Local Realism: An Analysis of Social Choice Theory

Obdulia Torres González

When referring to realism in science, physics immediately springs to mind, and there are a number of arguments that help to uphold a realist position: the causal relationships established through assays, as defined by I. Hacking (Hacking 1983), the miracle argument formulated by H. Putnam (Putnam 1978) or the empirical success over the course of time postulated by R. Boyd (Boyd 1973). Unfortunately, none of these arguments can be applied to social sciences. Generally speaking, and for authors such as Hacking (Hacking 1999, pp. 31–32), interactive classes constitute a fundamental difference regarding the matter of realism, in the sense that classifications in social sciences are interactive, which is not the case in natural sciences. This means there is a conscious interaction in social sciences between classes and individuals in terms of kinds of classification. Uskali Mäki also identifies differences, in this case involving the fields of economics and physics.[1] A series of questions therefore arise, such as: is it possible to be realist regarding quarks and leptons, and anti-realist regarding social structures and economic agents? Does the issue of realism have to be an all-encompassing matter, regarding science in general, or should we differentiate between social sciences and natural sciences? Further still, should differences be made regarding the different fields, and therefore refer to a realism in physics, a realism in economics, another in chemistry and yet another

[1]The four basic differences whereby the defence of realism in terms of physics is inapplicable to economics are as follows: the non-philosophical usage of the term realism among economists, the fact the ontology of economic theories does not adjust to the independent existence of the mind that is typical of the formulations of scientific realism, the question of theoretical terms and the impossibility of proving existence and truth in economics by invoking manipulability or success (Mäki 1996).

O.T. González (✉)
Departamento de Filosofía, Lógica y Estética, University of Salamanca, Campus de Unamuno, Edificio FES, 37007 Salamanca, Spain
e-mail: omtorres@usal.es

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_23,

one in sociology? Is it the same to refer to realism when we are dealing with mature theories, such as the general theory of relativity, as when we are dealing with new theories or one less confirmed, such as the rope theory? What does this imply for social sciences?

Mäki refers to a contextualised and local realism; that is, the question of realism is in fact the question of the realism of a theory and not a question about science in general, and it is in this sense that Mäki posits a local and contextualised realism (Mäki 1992a, 1996, 1998, 2005). The same stance is adopted by, among others, H. Kincaid, for whom "the realism issue in philosophy of science cannot be decided in a perfectly general way that ignores specific issues in specific science" (Kincaid 2000, p. 667). This is the strategy to be followed in this paper, addressing realism from the perspective of a specific theory, namely, social choice theory (SCT). The question underpinning this paper is whether one should hold realism about SCT in the sense posited by Mäki: "By a 'realist reading' of a theory I mean an interpretation of the theory as putatively referring to entities, such that the theory has a chance of being either true, close to the truth, or carrying the promise of getting us closer to the truth of what it represents" (Mäki 1992a, p. 38).

This will enable us to clarify certain important issues regarding the theory's theoretical state and its explanatory capacity.

The first step involves a brief characterisation of SCT, paying special attention to the contribution the philosophy of science may make to the theory. This is followed by an illustration of some of the issues the theory addresses, and finally, the question of realism is considered from a contextualised perspective.

Social choice theory stands at the crossroads between economics, politics and ethics, being all seasoned with a smattering of formal logic. In its most simple interpretation, it is a theory that addresses the manners and procedures for taking collective decisions; that is, how we aggregate individual decisions to reach a collective one.[2] From this perspective, a formal study is conducted of the different voting rules and the requirements that have to be fulfilled by any method of aggregating preferences in order to be considered a sound method. This is where we encounter one of the theory's better known outcomes, namely, Arrow's Impossibility Theorem, which states that any decision method that satisfies certain minimum conditions of rationality (reflexivity, transitivity and completeness) and four more or less trivial ethical conditions,[3] taken from the majority voting rule, will provide an intransitive or dictatorial outcome. In short, the conditions are inconsistent with each other. This is the formal part of the theory. Generally speaking, it is an illustration

[2]The *Handbook of Social Choice and Welfare* defines the theory's domain as follows: "Social choice theory is concerned with the evaluation of alternative methods of collective decision-making, as well as with the logical foundations of welfare economics" (Arrow et al. 2002, p. 1).

[3]The conditions are the independence of irrelevant alternatives, a weak version of the Pareto principle, unrestricted domain and non-dictatorship.

of the use of the axiomatic method, which can be used not only to define a voting mechanism, but also for the definition of any mechanism for collective decision-making, especially in distributive matters.

The economic part has more to do with individual models and the theory's mathematical tools. There is one field that has been characterised by an economistic approach since its inception, and that is the field of social choice theory. Indeed, Arrow himself singles out market and vote as two special cases of the more general category of collective social choice (Arrow 1963, p. 5). We may therefore refer to the ordering of preference, indifference curves, individual utility and social utility, alluding to *homo economicus* as the protagonist of these decisions that are to be aggregated, etc. The key question here is that the two theories, consumer behaviour theory and rational choice theory, share an essential part of their ontology. Furthermore, the bulk of these decisions that are addressed are decisions of a distributive nature. Strictly speaking, the alternatives presented in social choices are social states that are defined as:

> A complete description of the amount of each type of commodity in the hands of each individual, the amount of labor to be supplied by each individual, the amount of each productive resource invested in each type of productive activity, and the amount of various types of collective activities, such as municipal services, diplomacy and its continuation by other means, and the erection of statues to famous men (Arrow 1963, p. 17).

It stands to reason that when ordering the alternatives, or social states, at both an individual and collective level, criteria are called for to indicate which alternatives are socially preferable. These criteria, whether they involve equality, merit, needs, utility, etc., are ethical criteria, yet the theory does not distinguish between them. "A member of Veblen's leisure class might order the states solely on the criterion of his relative income standing in each; a believer in the equality of man might order them in accordance with some measure of income equality" (*ibid.*).

Finally, politics envelops everything insofar as it is defined as decision-making for achieving a group's goals in public affairs.

Choice theory is, or should be, a vast interdisciplinary field, where economists, politicians, moral philosophers and mathematicians come together, being furthermore a field that provides philosophers of science with an extremely fertile field of work to which they can apply a different treatment. Unfortunately, collaboration is scarce, and within this field philosophers of science are considered outsiders. My interest, in what follows, is to highlight some of the problems that may be addressed from the philosophy of science in this field of work and whose discussion will ultimately lead to the consideration of realism within social choice theory. These questions will also serve to illustrate how this theory works, over and above well reported results such as Arrow's Impossibility Theorem.

The first of these questions involves conceptual definitions and how these are applied in the theory. We shall begin by studying aspects of the axiomatic method. Generally speaking, the axiomatic method may be defined[4] by the design of a

[4]For a characterisation of the axiomatic method, see García Bermejo (2002).

functional mathematical figure that depicts, schematises or models the aggregation and decision models. Once this figure has been designed, the analysis goes on to focus on the sets of properties, or axioms, that these processes may, or should ideally, satisfy. There are sundry ways of applying the axiomatic method; first, there are the impossibility outcomes, of which Arrow's theorem is the paradigmatic example, where it is denied there is any function of a specific class that satisfies a specific set of conditions. A second approach would be to investigate the properties of, or the conditions fulfilled by, a specific defined function, and third, given a set of properties, find out which functions fulfil them. In each case, the mathematical function seeks to order the different social states according to their social desirability, with this concept being defined through the conditions the function has to satisfy. This latter point is the key one to be stressed; in other words, the ordering criterion is defined through the function's conditions. It is not a prior criterion but rather one that is built based on such conditions; what we want to say by this. It may be expedient to develop it through an example related to egalitarian distributions. The example presented forthwith is by B. Tungodden (2003). The author defines moderate egalitarianism (the criterion for ordering social states) as the combination of an interest in fostering equality with the principle of personal good. The conditions defining it are as follows: anonymity, conditional contracting extremes, principle of personal good and the strict priority to equality promotion. There now follows a definition of these conditions.

- *Anonymity*: For all alternatives x and y, if x is a permutation of the values of y, then x is equally as good as y.
- *Strong Conditional Contracting Extremes (or betterness)*: For all alternatives x and y, if (1) all the best-off persons in x are best-off persons in y and their well-being level is strictly lower in x than y; (2) all the worst-off persons in x are worst-off persons in y and their well-being level is strictly higher in x than y, and (3) the well-being of everyone else is the same in x and y; then x is better than y.
- *Strict Priority to Equality Promotion*: For all alternatives x and y, if (1) there are persons with higher well-being in x than y and persons with higher well-being in y than x, and (2) x is more equal than y, then x is better than y.
- *Principle of personal good*: For all alternatives x and y, if everybody is as well off in x as in y, and someone is strictly better off (in x) then x is better than y.

The problem is that the conditions described generate an intransitive result. Let us now consider an example. "Suppose that $y = (1, 100, 100)$ is considered more equal than $x = (2, 10, 100)$, and hence strict priority to equality promotion implies that y is better than x. Compare x with $z = (2, 10, 10)$. From the principle of personal good, it follows that x is better than z. By transitivity, we now have that y is better than z. However, this violates strict priority to equality promotion according to the minimal requirement of strong conditional contracting extremes on equality" (Tungodden 2003, p. 14). The option is therefore either to modify any one of the conditions or renounce transitivity. Tungodden's solution is to modify the conditions, defining equality according to the maximin principle, which selects the

situation that maximises the state of the worst-off person. This allows affirming that *x* is better than *y*, according to the definition of equality, and avoiding the intransitivity of the outcome.

This means the formal requisite of transitivity informs the concept's characteristics that are chosen to order the different social states. Not just any characterisation of equality will be valid, but instead only those generating quasi-orderings. For some philosophers, on the other hand,[5] the concept of equality does not generate transitive orderings, so this structure, and therefore the axiomatic method, needs to be renounced for ordering social states. What is at stake here is that a technical requisite informs the concept's characteristics that may or may not be chosen. This is one of the reasons for the clash between the advocates of the axiomatic method and its detractors, or, as some have put it, between economists and philosophers.[6] In simple terms, philosophers have an intuitive understanding of the nature of inequality, with their definition being embodied in principles that very often generate impossibility outcomes. Economists, for their part, define the concept according to the ordering generated by some or other set of principles. The concept is therefore designed or constructed by theoreticians through the measurement process.[7] What is ultimately at stake is the explanation of the conditions that order social states, and when it is undertaken in economic terms it is related to logical compatibility, whereas for philosophers the ordering should be decided by their notion of equality, regardless of possible inconsistencies or the possibility or not of ordering all the social states.

A further issue dividing economists and philosophers refers to the possibility of using experimentation as regards distributive theories.

The traditional role assigned to assays according to the standard conception is to test theories. This implies, as has been sustained, that experimentation has no relevance in the context of the theories of distributive justice, given that these are normative theories based on principles or ethical judgements. Since Hume's famous *dictum*, regarding the impossibility of inferring is-ought judgements, theoretical knowledge is generated by scholarly introspection and deduction from the general principles so founded. The general opinion is that popular beliefs about justice are one thing and a correct theory about it is quite another. The data obtained through empirical research are not relevant for the correction or validity of the theories. However, in recent years there has been a growing interest in research in this area, which seems to indicate that this view is not widespread and experimentation does have a role to play in this field.

I have posited elsewhere (Torres 2010) different roles for experimentation as regards distributive theories, so I shall focus here solely on the part it plays in

[5] See, for example, Larry Temkin especially (Temkin 1987, 1996).

[6] See (McKerlie 2003) and (Tungodden 2003).

[7] The same occurs with the necessary requisite of additive separability when we introduce the principle of personal good. Egalitarian distributions do not comply with separability, but when they do, it is at the price of renouncing equality.

conceptual clarification, in the sense it allows a conceptual exploration of the consequences that may be forthcoming from the definitions of the concepts.

One of the most important fields of research in egalitarian theories is the definition of the measures of inequality. The concept is highly complex, and a wide range of measures have been suggested for ordering social states in different ways. The research conducted by Yoran Amiel and Frank Cowell in 1991 (Amiel and Cowell 1992)[8] sought to prove experimentally whether the suggested axioms, as properties of the Lorenz curve, fitted the definition of the concept of inequality that individuals hold. In short, the Lorenz curve, one of the more popular mechanisms among economists for evaluating the inequality of different income distributions, is defined by the following axioms:

- *Anonymity*: for all the alternatives x and y, if x is a permutation of the values of y, then x is equally as good as y.[9]
- *The principle of population*: Replication of the population and its income should not affect the index of inequality.
- *The principle of scale invariance*: The only important thing is the relative benefit and not the absolute one. This means the index is not sensitive to proportional increments of the relevant benefit.
- *The Pigou-Dalton transfer principle*: Inequality always decreases when there are transfers from better-off individuals to worse-off ones, as long as the mean income does not decrease and the order among the individuals involved remains the same.

It should be noted that accepting the Lorenz curve means accepting every one of these axioms. The first three may be more or less questionable as they involve a choice between absolute measures and relative measures of inequality. The same does not occur with the Pigou-Dalton principle, which is almost unanimously accepted by theorists. By way of example: let us assume the following distributions: A = (1, 4, 7, 10, 13) and B = (2, 4, 7, 10, 12). According to the principle, inequality decreases in the step from A to B given that, first, there has been a redistribution from better-off to worse-off; second, the mean income has not changed, and third, the order among the individuals involved remains the same. Let us now consider the following distributions: A = (1, 4, 7, 10, 13), B = (1, 5, 6, 10, 13). If you consider that A is more unequal than B, this means you are a proponent of the Pigou-Dalton principle, whereas if you consider B is more unequal, then you do not support said principle. The fact this principle is not supported in this specific example may be due to two reasons. First, the transfer does not take place between individuals that we may associate to the situation of better- and worse-off, as they occur in the middle of the vector. Assays appear to show that agreement with the Pigou-Dalton condition depends on who transfers and who receives. When the transferor is the best placed

[8]The questionnaire is reprinted in full in the appendix to the paper.

[9]Even though anonymity may seem trivial, it is incompatible with the different rights individuals may have.

individual and the receiver the worse placed one, the principle receives high support, whereas when the transfers occur at the top of the table, that is, between better placed individuals, support for the principle decreases (Amiel and Cowell 1992, pp. 15–17). On the other hand, it can be affirmed that with the transfer, although individual 2 sees an increase in their income, this implies the distance between the third worse-off and the fourth has increased, and this may be seen as an increase in inequality.

What is the significance of these findings for the concept of inequality? Should theorists take them into account in their definition of the concept of inequality? Should the transfer principle be modified so that it accepts only transfers between the better- and worse-off, while rejecting transfers in the lower part of the table? Or are theorists in a privileged position for evaluating inequality? Can we associate the empirical validity of the theory to the extent to which it reflects the understanding the majority of individuals have of equality or justice? Accordingly, can an individual's moral intuitions play the same role as observation in science; in other words, can they help us to verify the theory? If that is the case, can that provide the pillars for upholding moral realism? In one sense, the moral intuitions of individuals we *discover* through experimentation seem to support a certain moral realism, in another, given their diversity they seem to reflect the particular moral theories dependent on their tradition or culture, which would support an antirealist position.

The question of moral realism is beyond this paper's scope, but to the extent there is an ethical dimension involved there are certain points to be made. Following R. Boyd, we understand moral realism to be the doctrine that maintains the following:

> Moral statements are the sorts of statements which are (or which express propositions which are) true or false (or approximately true, largely false, etc.): The truth or falsity (approximate truth . . .) of moral statements is largely independent of our moral opinions, theories, etc.; Ordinary canons of moral reasoning – together with ordinary canons of scientific and everyday factual reasoning- constitute, under many circumstances at least, a reliable method for obtaining and improving (approximate) moral knowledge (Boyd 1988, p. 307).

Boyd's strategy is to show that moral methods and beliefs are much closer to our scientific methods and beliefs, in the sense of external, empirical, inter-subjective and objective, than we are inclined to believe. In this sense, Boyd gives moral intuitions the same role as intuitive judgements in science that individuals learn through study and training in the discipline.

> But it seems overwhelmingly likely that scientific intuitions should be thought of as trained judgements which resemble perceptual judgements in not involving (or at least not being fully accounted for by) explicit inferences, but which resemble explicit inferences in science in depending for their reliability upon the relevant approximate truth of the explicit theories which help to determine them (Boyd 1988, p. 319).

The problem, as we saw earlier, is that economists and philosophers not only have different academic backgrounds but also different intuitions regarding what equality is, for example.

In view of the problems selected and the drift taken by the preceding paragraphs, the reader will have surmised we are heading towards a discussion of the theory's

realism. This, too, is a field that confronts philosophers and economists because of the completely different use these two groups make of the term realism. Let us once again cite Mäki:

> Economists use the term for the purpose of attributing properties to their representations, such as models and their assumptions. (. . .) In contrast, philosophers use the term realism to denote various philosophical theses or theories, such as existence (. . .) relations between the words and the world (. . .) justified knowledge claims (. . .) the goals of science (. . .) and so on (Mäki 1998, p. 301).

We are facing a three-pronged difficulty. First, the varied definitions of realism and the different spheres of application, whereby realism has different classifications. Second, the theory's level of abstraction and the importance of the formal component. Finally, the theory interweaves economic, political, moral and mathematical components.

The issue of realism becomes seriously complicated when an attempt is made to find a definition of realism or an appropriate classification. A differentiation should be made between ontological, epistemological, semantic, theoretical or progressive realism (Dieguez 2005), or there is one realism, but we should distinguish between the metaphysical, epistemic and semantic approach, as propounded by S. Psillos (Psillos 2003, p. 60), D. Hausman, in turn, formulates realism on the basis of four postulates:

> Goals: science aims to discover the truth about its subject matter as well as to assist human practice . . . Truth: the claims theories make, including the claims involving unobservable, are true or false and should be true. Existence: the unobservable entities referred to by true theories exist. Knowledge: It is possible to have good reason or evidence for scientific theories, including theories that talk about unobservables (Hausman 1998, p. 191).

And Mäki distinguishes between an ontological realism, a referential one, a representational one and a veristic one. The crux of the matter is that the manner in which realism is defined will provide one response or another to the differences between natural sciences and social sciences. For example, if we consider the question of realism by posing the question: "is there a domain of objects that is independent of the mind – individual or collective – that can be expressed properly through language and reliably captured by knowledge?" (Gonzalez 1993, p. 14) The answer on the realism of the theory will depend on the definition provided on what it means that objects are independent of the human mind, although if we take that independence in the true sense of the word, there appears to be no place for realism in social sciences.[10] In this paper, when considering the realism of SCT, it is affirmed that the objects the theory refers to exist and the properties attributed to them are true or tentatively true, although our interest is focused mainly on ontological realism.

[10]U. Mäki propounds a notion of existence other than the notion of the independent existence of the mind that is typical of realism. This would be the notion of objective existence defined as: "X exists objectively relative to a given representation if it exists unconstituted by that particular representation (both material, social, and mental entities may exist objectively)" (Mäki 1996, p. 433).

The introduction to this paper indicated that social choice theory includes aspects of human choice that interweave economics, politics and ethics, and that the theory shares part of its ontology with consumer behaviour theory. The first issue regarding the realism of the theory being considered here is whether the question of realism should apply to all these aspects, or refer instead solely and exclusively to human choices and their component parts; that is, desires and beliefs or preferences and expectations. Our first question, therefore, is: what is the ontology of the theory?

If we focus solely on the economic sphere, we find that the ontology of the theory is exactly the same as that we would postulate for microeconomic theory, which in the case that concerns us would be limited to people's system of choice and the properties advocated for it; basically, rationality as consistency and the maximisation of utility as selection criterion. Hausman provides a realistic response to the problem in hand. According to Hausman, subjective probabilities and preferences are idealised variations of the notions of desires and beliefs with which we function in everyday life. It is therefore difficult to affirm that all other human beings have desires and beliefs, yet deny there are systems for ordering subjective probabilities or preferences (Hausman 1998, p. 199). Along these very lines Mäki affirms:

> the existence of the objects of the scientific realm should not be a major issue in economics. The referential realisticness of the fundamental elements of economic theories is more often than not beyond doubt: since the terms of economic theories seem to refer to entities with which economists and others are familiar on the basis of ordinary experience (Mäki 1996, p. 434).

We can, therefore, provisionally provide a realistic answer to the question of what Mäki refers to as the "ontic furniture" of our theory, in the sense that the objects the theory refers to do exist, even if it is only commonsense realism as defined by the authors. There are two important issues here: idealisation, mentioned both by Hausman and by Mäki, and the implications of commonsense realism. Commonsense realism means realism regarding the objects and representations of common sense, such as green cucumbers and fat politicians. This realism is different to the scientific realism that involves realism regarding scientific objects and their representations (Mäki 1992b, p. 174). In our analysis, in keeping with these scholars, we would have a commonsense realism regarding the behaviour of the actors that is conceptualised in terms of folk psychology, that is, as intentional actors with beliefs and desires and whose actions are governed by them. This characterisation is replaced in the theory by the rational agent, with an ordering of preferences that is reflexive, transitive and complete, with perfect information and capable of maximising a utility function. This replacement occurs through abstraction and idealisation (Mäki 1998). In fact, staying with Mäki, there are two ways in which theoretical representations stray away from commonsense representations, namely, modification and reordering. Modification means the reformulation of language expressions in terms of technical vocabulary and a formal register. Other more

substantial modifications include selection, abstraction, idealisation, exaggeration,[11] projection, aggregation and their multiple combinations (Mäki 1996, pp. 433–436). We have referred to the minimum version of commonsense realism, the radical version means realism regarding the objects of common sense and its representations and antirealism regarding scientific objects and their representations (Mäki 1992b, p. 174). The minimum realism of common sense would affirm that individuals have beliefs and desires, while radical realism would say there are no reflexive, transitive and complete orders of preference, nor individuals that maximise utility. This may have two readings: the one posited by Mäki, when he affirms that an economist's stance would be to consider that the supposed events involved in the agent's behaviour are unrealistic[12] and a second reading that simply involves being antirealist regarding the rational choice theory.

> The preferences of consumers are represented by the axioms of standard neoclassical theory as complete (. . .), reflexive (. . .), and transitive (. . .). It would be a mistake to conclude that if consumers do not have preferences with these characteristics, they do not have preferences at all. There may be other reasons to doubt the reality of preferences (along with the rest of the folk psychological realm), but, say, the intransitivity of preferences should not be one such reason. The axioms of consumer theory may refer to real entities irrespective of how these entities are represented (Mäki 1996, p. 436).

We have established the possibility of being realists, at least a minimum ontological realism, regarding the theory's economic sphere, but what occurs as regards the political component or the ethical component? What is the "political furniture" of our theory? There are voting rules, decision-making committees, political options to decide between, distribution rules, political constitutions, Pareto optimal outcomes . . . On a trivial level, all these items are constructed, in the sense that if human beings did not exist they would not exist either, but that does not stop them being real. A useful distinction along these lines is the one that I. Hacking takes from J. Searle among items that are ontologically subjective but epistemologically objective (Hacking 1999, p. 22).

What happens with the theory's moral component? Although the two cases being analysed have a constructivist flavour, care needs to be taken when referring to realism regarding social choice theory. The notions of equality or inequality analysed here pertain to the sphere of people's beliefs regarding what is, socially speaking, a fair or preferable state. Regarding ontological realism, there is no difference between an individual ordering alternatives according to a criterion of personal interest and ordering them according to their equality, or by those that are socially fairer. As we have already noted, the theory is neutral regarding the values used in the ordering, and what's more, this is the true performance of the theory's characteristic axiomatic method, but R. Hardin has made a point that is relevant here. We are referring to the ordering of social states by individuals

[11] According to the author, the assumption of maximisation and transitive preferences is an exaggeration.

[12] In the sense of not being a true picture of reality.

according to the relevant ordering criterion; that is, we are returning once again to the model of rational choice that underpins the theory. According to Hardin, the theory's neutrality regarding the underlying ordering value converts the theory into a conceptual instrument rather than an explanatory one. Whereas the theory is a suitable explanatory instrument when it is based on individual self-interest, it is merely an exploratory tool when it lacks that value or is neutral among values, given that Arrow's result is forthcoming regardless of the value used to order social states. Hardin affirms:

> To a large extent, unless we can carry out a program such as Becker's, the difference between assuming some value theory for the actors and not assuming any is the difference between explanatory and conceptual analysis in individualistic political economy. If we impute certain substantive values to the actors in our theory, we can say what behaviors or choices follow from those value commitments, If we do not, we cannot say much (Hardin 2001, p. 68).

Valorative neutrality appears to contradict the fact that social choice theory and consumer behaviour theory share part of their ontology, that is, rationality as consistency and the maximisation of utility as an ordering criterion, with it all depending on how we define the utility. The contradiction arises if the maximisation of utility is interpreted as self-interest, although in theory utility does not imply self-interest.

We should at this point assess the possibility of making a realistic reading of SCT, and we have to admit that it is difficult to maintain a realistic stance towards it. The most we have managed to establish is a commonsense realism regarding the agents that, in its radical version, is compatible with a scientific antirealism or with a heavily idealised and abstract version of the theory. The question is that with the theory characterised in these terms, the analysis also concludes in an antirealist position. Let us see why.

B. Ellis distinguishes between different types of explanations and theories as a relevant issue in the question about realism (Ellis 1985).[13] We therefore have causal, functional, model and systemic theories as well as causal, functional, model and systemic explanation.

> A model theoretic explanation is information about how (if at all) the actual behavior of some system differs from that which it should have ideally if it were not for some perturbing influences may be causing the difference. (. . .) Model theories define norms of behaviour against which actual behaviour may be compared and (causally) explained (Ellis 1985, p. 55).

It seems fairly clear, as we have been affirming throughout this paper, that SCT is a model theory and the type of explanations available to us are theoretical model explanations. The most important question is, according to Ellis, that realism is applied to causal theories and not to theoretical models. This is because postulating that A causes B is to accept that both A and B exist. Nevertheless, the hypothetical

[13] I have taken the idea of using Ellis' theory from U. Mäki (1992a), who uses it in his analysis of Austrian theory.

entities of the model theories are not the cause of anything and so there is no reason to believe that the entities they refer to exist. This would be applicable not only to the core of the theory related to the choice of the agents[14] but also to the distributive mechanisms described here. These mechanisms explore the outcomes that would be forthcoming in a distributive sense applying the different notions of equality and inequality. A theory of these characteristics has no place for realism. Therefore, the question of the theory's realism is a question about its theoretical status and its explanatory capacity. This means the theory is not explanatory, it does not shed light on the agents' true choices, nor does it predict the results of those choices.

Both the neutrality in terms of values and the idealisation of the theory head in the direction indicated by Ellis, the theory is not explanatory, it is a useful tool, a mechanism for conceptual exploration, but its entities do not exist nor are its propositions susceptible to be true or false given their ideal nature. The conclusion, therefore, is that it is not possible to make a realistic reading of SCT, which does not imply, nonetheless, that there is no realist positioning regarding other social theories, with their status regarding realism being something that will have to be clarified on a case-by-case basis.

Acknowledgement This work has been possible thanks to the funding provided by Spanish Ministry of Science and Innovation (FFI2012-33998).

References

Amiel, Y., and F. Cowell. 1992. Measurement on income inequality, experimental test by questionnaire. *Journal of Public Economics* 47: 3–26.

Arrow, K.J. 1963. *Social choice and individual values*. New York: Wiley.

Arrow, K.J., A. Sen, and K. Suzumura. 2002. *Handbook of social choice and welfare*. Amsterdam: Elsevier.

Boyd, R. 1973. Realism, underdetermination, and a causal theory of evidence. *Noûs* 7: 1–12.

Boyd, R. 1988. How to be a moral realist. In *Moral realism*, ed. G. Sayre-McCord, 307–356. New York: Cornell University Press.

Dieguez, A. 2005. *Filosofía de la Ciencia*. Madrid: Biblioteca Nueva.

Ellis, B. 1985. What science aims to do. In *Images of science*, ed. P. Churchland and C. Hooker, 48–74. Chicago: Chicago University Press.

García Bermejo, J.C. 2002. Sobre el método axiomático en la teoría de la elección social. In *Enfoques filosófico metodológicos en Economía*, ed. W.J. Gonzalez et al., 217–274. Madrid: FCE.

Gonzalez, W.J. 1993. El realismo y sus variedades: El debate actual sobre las bases filosóficas de la Ciencia. In *Conocimiento, Ciencia y Realidad*, ed. A. Carreras, 11–58. Zaragoza: Mira Editores.

Hacking, I. 1983. *Representing and intervening*. Cambridge: Cambridge University Press.

Hacking, I. 1999. *The social construction of what?* Cambridge, MA: Harvard University Press.

[14]This is not a new idea, as many consider the theory of rational choice that underpins microeconomic theory and SCT to be a normative theory.

Hardin, R. 2001. The normative core of rational choice theory. In *The economic world view. Studies in the ontology of economics*, ed. U. Mäki, 57–74. Cambridge: Cambridge University Press.

Hausman, D. 1998. Problems with realism in economics. *Economics and Philosophy* 14: 185–213.

Kincaid, H. 2000. Global arguments and local realism about the social sciences. *Philosophy of science* 67 (Supplement. Proceedings of the 1998 Biennial meetings of the Philosophy of Science Association): S667–S678. Published by The University of Chicago Press.

Mäki, U. 1992a. The market as an isolated causal process: A metaphysical ground for realism. In *Austrian economics: Tensions and new directions*, ed. J. Davis et al., 35–59. New York: Kluwer Academic.

Mäki, U. 1992b. Friedman and realism. *Research in the History of Economic Thought and Methodology* 10: 171–195.

Mäki, U. 1996. Scientific realism and some peculiarities of economics. In *Realism and anti-realism in the philosophy of science*, ed. R.S. Cohen et al., 427–447. New York: Kluwer Academic.

Mäki, U. 1998. Aspects of realism about economics. *Theoria* 13(2): 301–319.

Mäki, U. 2005. Reglobalizing realism by going local, or (how) should our formulations of scientific realism be informed about the sciences? *Erkenntnis* 63: 231–251.

McKerlie, D. 2003. Understanding egalitarianism. *Economics and Philosophy* 19: 45–60.

Psillos, S. 2003. The present state of the scientific realism debate. In *Philosophy of science today*, ed. P. Clark and K. Hawley, 59–82. Oxford: Oxford University Press.

Putnam, H. 1978. *Meaning and the moral science*. London: Routledge.

Temkin, L. 1987. Intransitivity and the mere addition paradox. *Philosophy and Public Affairs* 16: 138–187.

Temkin, L. 1996. A continuum argument for intransitivity. *Philosophy and Public Affairs* 25: 175–210.

Torres, O. 2010. The role of experiments in the theories of distributive justice. In *New methodological perspectives on observation and experimentation in science*, ed. W.J. Gonzalez, 159–169. A Coruña: Netbiblo.

Tungodden, B. 2003. The value of equality. *Economics and Philosophy* 19: 1–44.

Objectivity and Visual Practices in Science and Art

Chiara Ambrosio

1 Objectivity and Its Histories

> All epistemology begins in fear – fear that the world is too labyrinthine to be threaded by reason; fear that the senses are too feeble and the intellect too frail; fear that memory fades, even between adjacent steps of a mathematical demonstration; fear that authority and convention blind; fear that God may keep secrets or demons deceive. Objectivity is a chapter in the history of intellectual fear, of errors anxiously anticipated and precautions taken (Daston and Galison 2007, p. 372).

Objectivity is a contentious concept. It divides and generates debate; it is used as a weapon and as a defence against uncertainty and inaccuracy; it placates the fear that the objects of scientific investigation are not as stable as we wish they would be. As Daston and Galison insightfully state, objectivity stands out as one of the most prominent chapters in the history of intellectual fear, and as such it can be both liberating and oppressing. Yet, it is exactly because of its contentious status that objectivity continues to surprise us. More importantly, it is because of its contentious status that we still have much to say about it.

My aim in this paper is not to define objectivity. Instead, and in line with Daston and Galison, I want to illustrate objectivity – and I use "illustrate" in quite a literal sense. Daston and Galison taught us that objectivity has a history, which is best told through pictures. The atlas images discussed in their still much debated book, *Objectivity,* revolve around the practice of image-making and the epistemic virtues – objectivity being one among many – that inform and guide the scientific quest for accurate representation. Scientific atlases, in their narrative, are a case in point: they

C. Ambrosio (✉)
Department of Science and Technology Studies, University College London, Gower Street, WC1E 6BT London, UK
e-mail: c.ambrosio@ucl.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_24,

offer a glimpse of how collective practices of observation and shared styles of visualization become essential to "discern and stabilise" (Daston 2008, p. 98) the objects of scientific investigation.

In the course of this paper, I add a new layer to Daston and Galison's narrative, by showing that scientists were not alone in this process, and that in many cases they did not have the final word on what count as accurate representations. My aim is to show that the history of scientific objectivity has constantly crossed paths with the history of artistic visualisation, from which it has received some important challenges. The story that I present is one about the contingencies that underpin what we usually regard as ready-made images, and the controversies that arise when artists and scientists respond to each other's modes of visualisation. I claim that the very nature of such controversies played a crucial role in the history of objectivity, and that philosophical accounts of objectivity have much to gain from a closer engagement with this history.

2 Idealisation and Its Discontents

Daston and Galison's narrative begins well before the birth of objectivity. In their account, truth-to-nature is the representative standard that eighteenth century savants pursued before the term "objectivity" became the hallmark of accurate representations: "[Eighteenth century] images were made to serve the ideal of truth – and often beauty along with truth – not that of objectivity, which did not yet exist" (Daston and Galison 2007, p. 104). Truth, perfection, beauty: these commitments were considered complementary to each other, rather than mutually exclusive, by eighteenth century idealisers. Truth-to-nature imposed familiarity with nature's variations, which served the purpose of taming variability. Eighteenth century savants were expert idealisers, and had the scientific duty to correct nature for the sake of truth. Their illustrations show that in this time scientific representation was continuous with idealisation, construed explicitly as the act of extracting the ideal form from individual instances.

The ideal of truth-to-nature was explicitly pursued through the collaboration with artists – a collaboration aimed at the fusion of the head of the scientist with the hand of the artist (*id.*, p. 88). Daston and Galison stress the regimes of discipline enforced by scientists, who regarded their illustrators as subordinate to the greater cause of scientific accuracy. The very nature of these collaborations imposed that the eye and hand of the artist should undergo strict training, with scientists often intervening on drawings and sketches to ensure their scientific correctness. But the story of artists and scientists working side by side is also one of conflict and controversies, of scientists enforcing ideas of perfection and truth and artists reacting, in more or less overt ways. The Leiden anatomist Bernhard Siegfried Albinus and his collaborator, the artist and engraver Jan Wandelaar, are a remarkable example of how controversy and disagreement affected the path of truth-to-nature.

What Albinus was after, in his monumental *Tabulae Sceleti et Musculorum Corporis Humani* (1747), was an accurate representation of the human body, one which could take him beyond nature's variety and individual flaws or defects, and condense his philosophical ideal of "homo perfectus" (Punt 1983; Hildebrand 2005, p. 557). In order to achieve this ideal, Albinus sought the help of the skilled Wandelaar. Strangely enough, historians (including Daston and Galison) have paid scant attention to Albinus' collaborator. Yet, his active intervention in the layout of Albinus' tables, which took the form of a silent rebellion against the impositions of scientific accuracy, is what ultimately determined their lasting impact on the general study of anatomy for over a century.

Albinus and Wandelaar had met in 1723, when they were both working on a re-edition of Vesalius' *De Humanis Corporis Fabrica* (Huisman 1992, p. 2; Hildebrand 2005, p. 559). What immediately brought the artist and anatomist together was their common concern about the imprecision of Vesalius' anatomical work, a concern that eventually led to their subsequent collaboration on a new – more informed and more precise – treatise on anatomy. Following the death of his son, Wandelaar moved into Albinus' house and lived there, continuing his training in anatomy, for over 20 years. It was during this time that the *Tabulae* were completed.

By the start of his work on the plates of Albinus' *Tabulae*, Wandelaar was already an accomplished illustrator and engraver. He had been collaborating with scientists at least since the 1720s, having designed and executed – among other things – the frontispiece of Linneus' *Hortus Cliffordianus*, published in 1737 (Daston and Galison 2007, p. 57). More importantly, with Wandelaar Albinus had found a convenient way to avoid one of the major sources of inaccuracy that had affected anatomical representations until his time: the mismatch between anatomical visualisation and its representation in the final engraved image. The passage from anatomical preparations to drawings, and from these to copperplates or woodblocks, required the intervention of an engraver to transfer the artist's drawings on plates. But while the artist's eye and hand had been trained directly by the anatomist, the engraver only entered this process in its final stages – and this was the cause of most mistakes. Wandelaar could offer both – drawing and engraving – thus securing the necessary continuity in the process of producing anatomical representations under the close supervision of Albinus himself.

Despite this, Albinus still demanded full control over his collaborator's work. A major concern for Albinus was the effect of distortion deriving from drawing in perspective. The problem was due to foreshortening, and to the fact that the artist could observe only part of its subject at a right angle. Thus, whatever was in the centre of the artist's field of vision, and viewed frontally, was depicted correctly. But the parts that were further removed from the centre were observed – and consequently reproduced – at progressively sharper angles, with the consequence of distortions and inaccuracies in the rendering of anatomical features. In an attempt to preserve both anatomical detail and correct representations, Albinus devised a two-steps method that constituted a genuine innovation in anatomical illustration. This involved a system of double grids, or "diopters", which would allow the artist to maintain the proportions of the "homo perfectus", along with preserving the most accurate degree of detail at the right angle of observation (Huisman 1992, p. 3).

Historians have paid considerable attention to Albinus' method[1] – and rightly so, as his experiments in measurement and accurate representation were among the first of their kind in the eighteenth century. There is, however, an aspect of the *Tabulae* that has remained relatively neglected in the literature, and that constitutes their most striking feature: the backgrounds of the engravings. Albinus' idealised bodies are placed in floral landscapes, surrounded by neo-classical architectural elements. The images and symbols of vitality presented in the backgrounds were in part related to Albinus' ideas about the unity of nature (Hildebrand 2005, p. 561); however, they also constituted Wandelaar's hard-gained space for artistic expression. It was in fact Wandelaar who convinced Albinus about the importance of the backgrounds of the plates, which he justified as an effort "to preserve the proper light of the picture, for if the space around the figure and between its parts were white, the light would suffer".[2] Alas, Wandelaar's effort to preserve "the proper light of the picture" resulted in the famous plate IV, which eventually became the symbol of the *Tabulae*, where Albinus' ideal skeleton is depicted with an equally well-proportioned rhinoceros in the background.

The animal in the background has been identified as Clara, a female Indian rhinoceros that arrived in Leiden in 1741 and travelled through Europe between 1746 and 1756 (Clarke 1974, p. 116; Rookmaaker 2005, p. 239). Historians explain the presence of Clara in the background of Albinus' illustrations as an exotic rarity which added an element of sophistication to the *Tabulae* (Hildebrand 2005, p. 561; Daston and Galison 2007, p. 72). This is partly true (Albinus himself substantiated this explanation in the text accompanying the plates), but there is more to this story. What historians often neglect is that, by the publication of Albinus' *Tabulae* in 1747, Wandelaar had been drawing images of the rhinoceros for at least 20 years. In 1727 he was commissioned to illustrate the Dutch translation of the complete description of the geography, ethnography and natural history of Cape of Good Hope written by the German naturalist Peter Kolb. The work, originally published in German in 1719, devoted almost two folio pages to a description of a two-horned African black rhinoceros. Interestingly, the plates of the German edition conflicted with Kolb's description, in that they portrayed the rhinoceros according to a tradition dating back to Dührer's 1515 iconic representation of the animal. Following Dührer, most of these illustrations, mainly based on second-hand accounts of the rhinoceros' features, presented the animal as one-horned (which is characteristic of the Indian species), covered by a thick armour and with a smaller spurious horn on the shoulders (Rookmaaker 2005, pp. 365 and ff).

Wandelaar's commission required him only to copy the plates from the 1719 edition of Kolb's book, but in fact he produced two different representations, both eventually included in the final publication: a traditional, Dührer-like depiction of the rhinoceros, which matched the illustration included in the German edition, and

[1] For a reconstruction of Albinus' diopters system see Huisman 1992, pp. 6 and ff.

[2] Quoted in Elkins (1986, p. 94), and Ferguson (1989, p. 232). Both report Albinus' quote (originally in the *Academicarum Annotationum*) from earlier editions of Choulant 1962.

a second image, which corrected and rectified it by following meticulously Kolb's description (Rookmaaker 1976, p. 88). In this second illustration, Wandelaar's black rhinoceros is represented, probably for the first time, with a smooth skin and two horns. Indeed, the captions to the illustration well capture the contrast that Wandelaar was trying to draw with his two conflicting images. The Dührer-like illustration is described as "The rhinoceros as it had been commonly depicted", whereas Wandelaar's rectified image is said to represent "The rhinoceros according to this description" (Kolb 1727, pp. 189–190). With this bold move, Wandelaar became one of the first artists who broke with the established tradition that had dominated the iconography of the rhinoceros – and even affected its classification – for over 200 years.

The story of Wandelaar's fascination with the rhinoceros and his role in the development of its iconography renders the presence of the animal in Albinus' *Tabulae* much less surprising. Far from being merely a fanciful and sophisticated addition to Albinus' anatomical works, the background of plate IV tells a story which runs in parallel to the one presented in the foreground of the engravings – a story about the insightful ways in which artists interfered with the criteria for accurate representation imposed by scientists to pursue their own agendas, in ways that eventually became even scientifically acceptable.

The inclusion of the rhinoceros in Albinus' plates involved a great deal of negotiation. The fact that Clara belonged to the one-horned, Indian species of rhinoceros helped Wandelaar's cause, as her features were far more compatible with traditional Dührer-like representations of the animal than the two-horned African species he drew in 1727. Hints of the reasons that Wandelaar might have used to persuade Albinus can be inferred from the commentary to the tables, where Albinus himself explicitly justifies the presence of the rhinoceros in the background of the plates as follows:

> We conclude this table, and the eight, by exhibiting in the back ground the figure of a female Rhinoceros that was shewed to us in the beginning of the year 1742, being two years and a half old, as the keepers reported. We thought the rarity of the beast would render these figures of it more agreeable than any other ornament resulting from mere fancy. The figures are just, and of a magnitude proportionable to the human figure contained in these two tables (Albinus [1747] 1749, sig.g.1.v.).

Here the presence of the rhinoceros in the plates is justified by the fact that her depiction is "just" and "proportionable" to the human skeleton. At the same time, however, Albinus confesses between the lines that the plates partly betray his ideal of truth-to-nature. What one sees in the background is a particular rhinoceros, observed alive in 1742, when she was two and a half years old. Wandelaar managed to sidestep Albinus' quest for idealised types, and brought the particular right at the core of the two most representative illustrations of his anatomical atlas. Clara's cumbersome presence in the background of the plates vindicates the role of Wandelaar, and with him the crucial contribution of artists, in shaping and challenging what counted as an accurate representation in the eighteenth century.

We can now go back to Daston and Galison's claim that eighteenth century truth-to-nature required the subordination of artists to scientists and read it in a slightly

different light. Some artists, like Wandelaar, approached accurate representation with all the dilemmas arising from the conflict between the needs of artistic experimentation and the impositions deriving from the canons of scientific accuracy. Hence, while the idealised bodies in the foreground of Albinus' illustrations are a sign of the ethos of discipline (well epitomised by the double grids and calculations) enforced by the scientist upon the artist, the rhinoceros in the background of plate IV shows that in some cases artists opposed their resistance against the visual restrictions imposed by the pursuit of truth-to-nature. Ultimately, the story of Wandelaar's rhinoceros shows that eighteenth century artists did not necessarily sacrifice their commitments for the cause of scientific accuracy: instead, they approached scientific illustrations with their own visual priorities, leaving more or less visible traces of their presence in the pictures.

The tensions, conflicts and visual arguments that characterise the relationship between artists and scientists add a new dimension to Daston and Galison's narrative, and they are not restricted merely to the eighteenth century. In the mid-nineteenth century a new representational mode, that of mechanically reproduced images, saw artists and scientists engaged in a different kind of battle around the status of accurate representations. This controversy was openly construed under the heading of "objectivity", a term which entered the artist's vocabulary when photography made its first appearance in artistic practice.

3 Mechanical Reproducibility and Its Discontents

Daston and Galison place the emergence of the modern concept of objectivity in the mid-nineteenth century, almost concomitantly with the birth of photography. Despite the fact that the concept extends to a broad range of recording instruments, photography constitutes a keystone in the process that led scientists to adopt an attitude that they define as "noninterventionism" toward the objects of their inquiries:

> One type of mechanical image, the photograph, became the emblem for all aspects of noninterventionist objectivity . . . This was not because the photograph was more obviously faithful to nature than handmade images – many paintings bore a closer resemblance to their subject matter than early photographs, if only because they used color – but because the camera apparently eliminated human agency (Daston and Galison 2007, p. 187).

Mechanical reproducibility imposed entirely new constraints on what counted as an accurate representation, and in this it contrasted sharply with the eighteenth century ideal of truth-to-nature. Eighteenth century representations required the intervention of the scientist – indeed, active intervention was just what conferred images credibility scientific reliability. Mechanical objectivity, on the other hand, required an attitude of asceticism toward the objects of scientific inquiry (*id.*, pp. 120 and ff). Letting nature speak for itself became the nineteenth century criterion for accurate representation, human intervention being now replaced by a procedural use of images which would ensure the removal of the scientist's judgment from

the process of image-making. This form of objectivity went hand in hand with an increased reliance on recording instruments, which, like the camera, promised the possibility of eliminating human agency altogether.

Contrary to scientists, artists sought in photography a new medium for aesthetic expression. Their quest for new ways of enhancing willful intervention led them to engage with the most technical aspects of the photographic process since the earliest stages of its development. In open conflict with scientific photography, artistic photography was conceived as a form of aesthetically-motivated resistance to the supposed "objective" status of the mechanically produced image, and this process saw artists treating the expressive possibilities offered by the camera as complementary and comparable to painting. Pictorialism, a movement that became dominant in the 1890s, explicitly pitted artistic photography against scientific photography by treating the former as painting. Pictorialist photographers accomplished this by selecting the content and the perspective from which photographs were taken, and intervened on the plates by directly retouching them. This practice brought the artist's subjective intervention right at the core of technical photography – indeed, it aimed to stress the impossibility of removing agency from photography, no matter what scientists thought or how they used their photographic equipment.

Mechanical objectivity saw artists overtly reacting to science, in a manner that contrasts rather sharply with the negotiations that characterized Wandelaar's intrusion in the background of Albinus' drawings. The controversy over objective images often took sarcastic tones, as in a 1903 article aptly entitled "Ye Fakers", in which the pictorialist photographer Edouard Steichen explicitly mocked the attitude of asceticism preached by the supporters of objectivity:

> Some day there may be invented a machine that needs but to be wound up and sent roaming o'er the hill and dale, through fields and meadows, by babbling brooks and shady woods – in short, a machine that will discriminatingly select its subject and by means of a skillful arrangement of springs and screws, compose its motif, expose the plate, develop, print, and even mount and frame the result of its excursion, so that there will remain nothing for us to do but send it to the Royal Photographic Society's exhibition and gratefully to receive the Royal Medal (Steichen 1903, p. 107).

Steichen's article first appeared in a journal whose mission was to advance artistic photography, granting its status as a form of art in its own right. The journal's name was *Camera Work*, and its founder, Alfred Stieglitz, occupies a central place in modernist responses to mechanical objectivity. Contrary to his pictorialist contemporaries, Stieglitz did not reject objectivity altogether. On the contrary, his works were eventually described by the members of his circle as the embodiment of objectivity in artistic practice. For instance, the art critic, caricaturist and amateur mathematician Marius de Zayas, describes Stieglitz's practice with tones that strongly remind of the rhetoric of noninterventionist objectivity:

> The desire of modern plastic expression has been to create for itself an objectivity. The task accomplished by Stieglitz's photography has been to make objectivity understood, for it has given it the true importance of a natural fact . . . Stieglitz, in America, through photography, has shown us, as far as it is possible, the objectivity of the outer world (De Zayas 1913, p. 13).

De Zayas' comment needs to be taken with a few caveats. While it is true that Stieglitz brought a scientifically-driven concept of objectivity right at the core of photographic practice, it must be stressed that he did so with the awareness that genuinely noninterventionist objectivity was an unattainable ideal. Instead, by proposing an experimentalist aesthetics based on a view of the artist as a *trained observer*, Stieglitz challenged both the resistance to mechanical reproducibility pursued by pictorialist photographers and the noninterventionist attitude cultivated by scientists in the mid-nineteenth century. His scientific background, often dismissed by art historians,[3] allowed him to become one of the most interesting voices in the history of objectivity.

Before plunging into artistic photography, a career that earned him the title of impresario of modern avant-garde, Stieglitz had the opportunity to be trained in the climate of experimentalism that characterized German science in the 1880s. In 1881, after having moved from America to Berlin, Stieglitz entered the Charlottenburg Polytechnic and began a degree in mechanical engineering. In the 1880s Berlin hosted a lively scientific community, which attracted the young Stieglitz since his early days at the Polytechnic. In parallel with his initial steps in the field of photography, he attended lectures by prominent figures such as the physicists Hermann von Helmholtz and Heinrich Hertz, the physiologist Emil DuBois-Reymond and the anthropologist and pathologist Rudolf Virchow (Kiefer 1991, pp. 61 and ff; Lowe 2002, p. 73). But the figure who influenced Stieglitz in the most dramatic way, eventually compelling him to switch from engineering to chemistry, was the chemist August Hofmann.

Hofmann is well known for his work on coal tar and his contribution to the development of aniline dyes, which laid the foundations of the German dye industry. A student of Justus Liebig at the University of Giessen, he had been a pioneer in the transition from analytic to synthetic organic chemistry. Hofmann adopted and extended Liebig's methodology, whose distinctive trait was the integration of teaching and research in the practical setting of the chemical laboratory.[4] Since his early years under Liebig's guidance, Hofmann had structured his laboratory as a research community, in which chemical knowledge was conveyed through practice. Most of the daily learning happened by observing and doing, whereas lectures provided a theoretical background for students who lacked prior chemical training. The concept that practice, far from being subordinate to theory, was constitutive of it, became especially important to Stieglitz. The scientific aesthetics underpinning his practice as a photographer revolved around the idea that photography and science shared the same experimental basis and that in both cases theoretical considerations emerged as generalizations from practical experience. When, in 1905, Stieglitz established the Little Gallery at 291–293 Fifth Avenue, in New York, he characterized it as his "experimental station" (De Zayas 1910, p. 47), and organized it as a laboratory that

[3]The only exception to these accounts is Kiefer (1991).

[4]On Liebig's laboratory see Holmes (1989) and Jackson (2009). On Hofmann's adoption of Liebig's model see Bentley (1970, 1972) and Meinel (1992).

followed Hofmann's (and Liebig's) model. Indeed, his breakthrough as the pioneer of modernist photography and as the impresario of avant-garde art in America consisted in adopting a Liebig-inspired model of laboratory conceived as a social space with its community, collective observational practices, shared representational conventions and tacit ways of conveying knowledge through action. Moreover, just as a scientific research community, Stieglitz and his laboratory group disseminated their findings through the journal *Camera Work* (published between 1903 and 1917), which became one of the most important instruments for the promotion of avant-garde in the twentieth century (Eversole 2005).

Stieglitz's chemical training under Hofmann prevented him from subscribing unconditionally to the widespread attitude of extreme interventionism that characterised pictorialist photography. While pictorialism still maintained a prominent place in *Camera Work* throughout the years of its publication, Stieglitz departed from it to embrace a more complex aesthetic position, which he identified as "straight photography". This new mode of visualization hinged on trained observation, which Stieglitz considered as the main route to achieve objectivity through experimental inquiry. Stieglitz's concept of the "seer" behind the camera explicitly appeals to a scientific view of trained eye, whose active judgment selects and interprets relevant aspects of a complex reality, and transposes them in a "true" photograph:

> It is high time that the stupidity and sham in pictorial photography be struck a solar plexus blow . . . Claims of art won't do. Let the photographer make a perfect photograph. And if he happens to be a lover of perfection and a seer, the resulting photograph will be straight and beautiful - a true photograph (Stieglitz 1910, in Adato 2001).

I argue that Stieglitz's approach to photography anticipates the transition from the asceticism of mechanical objectivity to the community-informed ethos of inquiry that Daston and Galison (2007, pp. 309 and ff) characterize as "trained judgment". Distinctive of twentieth-century image-making, trained judgment was a reaction to the constraints imposed by mechanical reproducibility. This new representational mode incorporated scientists' progressive awareness that trained observation, rather than the "blind sight" of mechanical objectivity, was the primary feature of scientific visualization. Such an ethos of inquiry, which built interpretation in the process of image-making without depriving photographs of their "straight" character, was just what Stieglitz had cultivated within the experimental setting of his galleries.

Most of Stieglitz' works exemplify the role of judgment in straight photography. I will discuss two specific cases, *The Terminal* (1893) and *Two Towers* (1913), as two illustrations of the development of Stieglitz's approach to trained vision.[5] Both photographs exemplify how Stieglitz, at various stages in his photographic career, purposefully weaved active judgment into the concept of straight photography.

[5]I am grateful to the Wesleyan University Library for allowing me to examine two original photogravures from *Camera Work* of *The Terminal* (Wesleyan Library A/N 1991.30.56) and *Two Towers* (Wesleyan Library A/N 1981.9.1).

The Terminal is one of Stieglitz's most famous photographs. It was taken in New York, just at the end of a severe snow blizzard in February 1893. The key feature of the image, which renders it particularly memorable, is the sharp contrast between the steaming horses at the terminal and the surroundings covered in snow. At the time, Stieglitz had not yet fully theorized his idea of straight photography. Despite the fact that the unmodified print was going to acquire a prominent role in his experimental aesthetics only years later, the idea of trained observation was already a central feature of his photographic practice. In an article he wrote only a few years later, exalting the virtues of the (then very innovative) hand camera with which *The Terminal* was taken, Stieglitz states:

> In order to obtain pictures by means of the hand camera it is well to choose your subject, regardless of figures, and carefully study the lines and lighting. After having determined upon these, watch the passing figures and await the moment in which everything is in balance; that is, satisfies your eye. This often means hours of patient waiting (Stieglitz 1897, p. 27).

Nature does not speak by itself, and the experienced eye of the photographer cannot and should not be removed from the photographic process. Moreover, "hours of patient waiting" are seen by Stieglitz as the necessary complement to the function of the trained eye. In line with Daston and Galison, it seems that the figure of the photographer outlined by Stieglitz is shaped around specific epistemic virtues. But while discipline and restraint characterize Daston and Galison's nineteenth century image-makers, in the case of Stieglitz training and perseverance are the qualities behind a successful photograph. Rather than deferring agency to the mechanical function of the machine, Stieglitz used his own scientific training to justify the presence of an expert eye behind the camera.

Judgment is also prominent in *Two Towers*, where the focus is on the combination of a man in a bowler hat captured between the railings leading to the entrance of a building and the branches of a tree covered in snow. This creates a powerful clash with the title of the photograph, which refers to the towers of Madison Square Gardens and the Metropolitan Life Building in the background. Once again, Stieglitz is proving that photography, far from offering a faithful reproduction of events, is about training the eye to see form and structure.[6] The idea of straight photography, of which *Two Towers* is a representative instance, condenses the key features of Stieglitz's experimental aesthetics. By challenging both naïve photographic realism and the simplistic appeal to subjectivity pursued by pictorialist photographers, Stieglitz stressed that photographic representation relies inevitably on the trained observer's active judgment in visualizing and selecting salient properties.

Stieglitz was well aware that photography, as every act of observation, is theory-dependent. And the theory that informs photography is in turn shaped by the needs and goals of the photographer, along with his tools, chemical equipment and laboratory practice. His scientific training provided him with a renewed awareness

[6]I discuss Stieglitz's ideas about form and structure with reference to his 1907 work, *The Steerage*, in Ambrosio (forthcoming).

of this aspect of photography and of the experimental process that guides the photographer through hours of patient waiting until he discerns the conditions for a "true" photograph. This characterization of Stieglitz's approach to photography suggests that his formulation of objectivity was not at all concerned with mechanical reproducibility. Far from preaching restraint and asceticism, Stieglitz recognised that objectivity required a trained "seer", and that the informed activity of trained observers constituted the crucial connection between artistic visualisation and mechanically reproduced images.

Historians and philosophers of science far too often tend to confine the changes that affected scientists' views about objectivity – especially with respect to mechanical reproducibility – only within science. I argue that the artistic responses to the image of objectivity enforced by scientists matter, if we want to draw a fuller and richer account of how this powerful concept plays out in practice and develops historically. This broader look at the visual arguments between artists and scientists discloses reactions and responses to objectivity that deserve closer examination, primarily because they *did* feed into the very history of the concept. On the one hand, there were pictorialist photographers who reacted to photographic objectivity with the compelling means of artistic experimentation and singled themselves out because of their obstinacy with the subjective aspects of artistic practice against the attitude of non-interventionism preached by scientists. On the other hand, there were scientifically-minded artists, such as Stieglitz, who sought for a compromise between resistance and restraint. In this process, science was a parameter that artists had constantly in mind: whether they adopted it or reacted to it, science informed their visual inquiries into the ways in which representations capture perspicuously some aspect of the world.

4 The Future of Objectivity

Daston and Galison's narrative ends with a snapshot of the current scientific attitude toward objectivity: the contemporary shift from "representation" to "presentation" (Daston and Galison 2007, pp. 382 and ff). The new scientific images, they claim, no longer "re-present". Instead, they function as unmediated "presentations", which fulfil the purpose of manipulating the real – and they do so in a aesthetically pleasing ways. This new representational mode opens a new chapter in the story of the relations between artistic and scientific visual practices: far from being disciplined and restrained, artists seem now invited to take a central place in the scientific enterprise, as they have the role of privileged instruments of scientific visualisation. This suggests that the time of controversy, of artists having to hide themselves in the background of anatomical engravings or having to engage in open arguments against the reliability of mechanically reproduced images, is finally over.

Unfortunately, most contemporary collaborations between artists and scientists, usually in the form of artist in residence programmes, are not so straightforward. Artists are often relegated to mere accessories in the toolkit of scientific visualisation

Fig. 1 Martin John Callanan *A Planetary Order (Global Terrestrial Cloud)*, 2009 (Courtesy of the artist)

and representation – a passive role that still requires their vision to be disciplined in some way. A parallel (and equally perplexing) assumption is that the artist's work is a mere means to add a visually pleasant dimension to scientific visualisation, which would magically render science more communicable.

My concern is that, by restricting artists to dispensable accessories, scientists or scientific institutions involved in interdisciplinary collaborations are depriving themselves of an indispensable ground for controversy that ultimately contributes to the growth of scientific knowledge. The historical case studies I have presented so far show that looking at the controversies that divided artists and scientists over what counts as accurate representation contributes to add an entirely new layer to the story of objectivity. My claim is that a normative lesson can be drawn from this history: artists *should* continue to pursue their critical mission – in fact they should take it very seriously, because scientific practice benefits from it. In this respect, I propose that the critical mission of art shares some of the features that characterize the role of philosophy of science: that of questioning and challenging assumptions and modes of working that would otherwise be taken for granted by scientific practitioners.

My final case-study exemplifies this point. I draw on a 2009 work by the artist Martin John Callanan, entitled *A Planetary Order (Global Terrestrial Cloud)* (Fig. 1). Callanan's artwork is a physical visualization of real-time "raw" scientific data. It captures at a glance one second of readings, taken on 2 February 2009 at exactly 0600 UTC, from all the six cloud-monitoring satellites overseen by NASA and the European Space Agency. The readings were transformed into a 3-D computer model, which was then 3-D printed at the Digital Manufacturing Centre at the UCL Bartlett Faculty of the Built Environment (Hamblyn and Callanan 2009, p. 67).

The clouds that cover the globe's surface, and that only suggest the presence of continents underneath, create a powerful perceptual shift: patterns of clouds that seem transient and mutable when viewed from earth form a coherent "planetary order" when seen from space. *A Planetary Order* gives a visible form to information

that would have otherwise remained in the form of silent quantitative readings, whose self-evidence is generally accepted with no reservations by scientists. As a representation, the globe is not so much – or at least not exclusively – about the exact position of patterns of cloud over the Earth's surface. The piece seems instead to raise more fundamental questions about what count as supposedly "raw" data, and the various ways in which such data can be visualized.

Callanan's work was part of a broader project, carried out in 2009 in collaboration with the writer Richard Hamblyn. The project involved them respectively as an artist and a writer in residence working in an interdisciplinary team of researchers at the UCL Environment Institute. Their collaboration aimed to use artistic visualization as an instrument to examine and criticise the scientific rhetoric of "data". Their results converged in a book, aptly entitled *Data Soliloquies* (2009), which features Callanan's *A Planetary Order* as one of its most iconic illustrations, and which draws on the visual narratives of climate change to underline more broadly "the extraordinary cultural fluidity of scientific data" (*id*., p. 13). The book has far reaching implications about the ways in which scientific data are communicated to the public, their often spectacular modes of representation, and the mechanisms of persuasion that are implicitly built in the display of large quantities of information. But a striking – albeit perhaps less obvious – aspect of the book is that it addresses, partly through Callanan's visual work, some important epistemological questions concerning the relation between data and their representation.

From the outset, Hamblyn and Callanan state that their aim is to "interrogate" data and bring to the fore the assumptions that far too often remain hidden behind their supposed self-evidence:

> Our title, 'Data Soliloquies' ... reflects the ways in which scientific graphs and images often have powerful stories to tell, carrying much in the way of overt and implied narrative content; but also that these stories or narratives are rarely interrupted or interrogated. They are information monologues – soliloquies – displayed more for their visual and rhetorical eloquence than for their complex (and usually hard-won) analytical content (Hamblyn and Callanan 2009, p. 14).

Data are collected and constructed according to specific acts of judgment, which form intricate narratives and stories behind their visual immediacy and power. Hamblyn and Callanan's message is that the quantity of data available does not justify their self-evidence: data visualization involves first and foremost an act of discernment, in which patterns are cautiously carved out of the statistical uncertainty that surrounds them.

Along with the shift from representation to presentation, which Daston and Galison identify as the latest frontier of correct depiction, the rhetoric of scientific objectivity has progressively turned to data and the statistical correlations between them as a guarantee of scientific reliability.[7] This new appeal to objectivity

[7]In a well known 2008 *Wired Magazine* article, for instance, Chris Anderson prophesized "the end of theories" as a result of the data deluge: "The new availability of huge amounts of data, along

hinges on quantity – a "data deluge" – without taking into account that its visual manifestations are themselves informed by judgment, discernment and choice. The task of the artist – well exemplified by Hamblyn and Callanan's work – consists in giving a visible form to such an act of discernment, and this places artistic visualization once again in the uncomfortable position of serving as a challenge to the parameters that define scientific objectivity. Untied from the discipline of truth-to-nature and the restraints of mechanical objectivity, artists can now vocalize their objections to objectivity in ways that can be immediately heard by scientists, and that can explicitly feed into their practice.

The kind of attitude toward objectivity – or at least toward the supposed objectivity of scientific data – that these new controversies will open is still difficult to anticipate. But the fact that artistic practice is now looking at data – a foundational and still little questioned assumption in scientific practice – is a promise that artists' challenges to scientific modes of visualization will not spare the fundamentals.

5 Conclusions

In charting the story of how artists participated in the conversations and controversies surrounding accurate representations, I stressed that the history of scientific objectivity has constantly crossed paths with the history of artistic visualisation. Whether glaringly displayed on the pages of books, atlases and journals, or hidden in the background of eighteenth century engravings, artists' reactions to objectivity shaped its history. I argued that the story of the often conflictual relation between artists and scientists has a great deal to teach to historians and philosophers alike. For one thing, it adds new interpretative layers to images that are too often taken for granted and classified strictly as "scientific", or dismissed as "merely artistic". Looking at the history of how certain artistic and scientific representations came to be the way they are reveals that scientific visualisation in all its forms is imbued with aesthetic commitments, and that artistic visualisation constantly capitalizes on – and responds to – scientific and technological innovation. This history shows that objectivity is neither historically nor philosophically straightforward, and that new interpretative layers can continually be added to a concept that is far too often taken as an immutable feature of the "world out there". If Daston and Galison showed us that objectivity has a history, I hope to have completed their account by giving a voice to the actors that remained in the background of some of the pictures they describe.

---

with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all". Anderson (2008), available at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (Last accessed 18/04/2012). For a recent perspective on the philosophical issues surrounding data-driven science, especially in biology and the biomedical sciences see Leonelli (2012).

But my aim in addressing artists' responses to objectivity is also a practical one. The history I have charted – albeit briefly – in this paper, shows that image-making is the privileged ground for controversies that ultimately feed into scientific knowledge and contribute to its growth. Stating normatively that artists working within scientific institutions *should* pursue their critical mission might be a controversial – if not radical – statement, and yet history teaches us that science *does* benefit from a critical engagement with artistic practice. The role of the philosopher, in this case, is specifically to address the epistemological issues arising from these collaborations, and this task will be carried out far more effectively with one eye on the past. The key role of modelling, visualization and imaging technology in contemporary scientific practice offers a unique opportunity for historians and philosophers to look beyond the boundaries of science and explore the key areas in which artistic and scientific practice have challenged and complemented each other, and will continue to do so as new forms of visualization and image-making arise. The story of objectivity is full of surprises, and it is up to historians and philosophers to decide what its next chapter will be.

References

Adato, P.M. 2001. *American masters – A. Stieglitz: The eloquent eye*. New York: Winstar TV and Video.

Albinus, B.S. [1747] 1749. *Tabulae Sceleti et Musculorum Corporis Humani*. Leiden: J. & H. Verbeck. Translated: *Tables of the skeleton and muscles of the human body*. London: John and Paul Knapton.

Ambrosio, C. forthcoming. Objectivity and representative practices across artistic and scientific visualization. In *Visualization in the age of computerization*, ed. A. Carusi, A.S. Hoel, T. Webmoor, and S. Woolgar. London: Routledge.

Anderson, C. 2008. The end of theory. *Wired Magazine* 16(7). Available at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory. Accessed 18 Apr 2012.

Bentley, J. 1970. The chemical department of the royal school of mines: Its origins and development under A.W. Hofmann. *Ambix* 17: 153–181.

Bentley, J. 1972. Hofmann's return to Germany from the Royal College of Chemistry. *Ambix* 19: 197–203.

Choulant, L. 1962. *History and bibliography of anatomic illustration*. Translated and annotated by Mortimer Frank. New York: Hafner.

Clarke, T.H. 1974. The iconography of the rhinoceros. Part II: The Leyden rhinoceros. *The Connoisseur* (February issue): 113–122.

Daston, L. 2008. On scientific observation. *ISIS* 99(1): 97–110.

Daston, L., and P. Galison. 2007. *Objectivity*. New York: Zone Books.

De Zayas, M. 1910. Photo secession notes. *Camera Work* 30: 47.

De Zayas, M. 1913. Photography and artistic photography. *Camera Work* 42(43): 13–14.

Elkins, J. 1986. Two conceptions of the human form: Bernhard Siegfried Albinus and Andreas Vesalius. *Artibus et Historiae* 7(14): 91–106.

Eversole, T. 2005. Alfred Stieglitz's camera work, and the early cultivation of American modernism. *Journal of American Studies of Turkey* 22: 5–18.

Ferguson, J.P. 1989. The skeleton and the rhinoceros. *Proceedings of the Royal College of Physicians of Edinburgh* 19(2): 231–232.

Hamblyn, R., and M.J. Callanan. 2009. *Data soliloquies*. London: The UCL Environment Institute.
Hildebrand, R. 2005. Attic perfection in anatomy: Bernhard Siegfried Albinus (1697–1770) and Samuel Thomas Soemmering (1755–1830). *Annals of Anatomy* 187: 555–573.
Holmes, F.L. 1989. The complementarity of teaching and research in Liebig's laboratory. *OSIRIS, 2nd series* 5: 121–164.
Huisman, T. 1992. Squares and diopters: The drawing system of a famous anatomical atlas. *Tractrix* 4: 1–11.
Jackson, C.M. 2009. *Analysis and synthesis in nineteenth century chemistry*. Ph.D. thesis, University of London.
Kiefer, G. 1991. *Alfred Stieglitz: Scientist, photographer and avatar of modernism, 1880–1913*. New York/London: Garland Publishing Inc.
Kolb, P. 1727. *Nauukeurige en uitvoerige beschryving van de Kaap de Goede Hoop* [Exact and extensive description of the Cape of Good Hope]. Amsterdam: B. Lakeman.
Leonelli, S. 2012. Making sense of data-driven research in the biological and the biomedical sciences. *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 43(1): 1–3.
Linneus, C. 1737. *Hortus Cliffordianus*. Amsterdam.
Lowe, S. 2002. *Alfred Stieglitz. A memoir/biography*, 2nd ed. Boston: Museum of Fine Art Publication.
Meinel, C. 1992. August Wilhelm Hofmann – "Reigning chemist-in-chief". *Angewandte Chemie* 31(10): 1265–1398.
Punt, H. 1983. *Bernard Siegfried Albinus (1697–1770), On 'human nature'. Anatomical and physiological ideas in eighteenth century Leyden*. Amsterdam: B.M. Israël.
Rookmaaker, L.C. 1976. An early engraving of the black rhinoceros (*Diceros Bicornis (L.)*) Made by Jan Wandelaar. *Journal of the Linnean Society* 8: 87–90.
Rookmaaker, L.C. 2005. Review of the European perception of the African rhinoceros. *Journal of the Zoological Society of London* 265: 365–376.
Steichen, E. [1903] 1997. "Ye Fakers", *Camera Works* (1): 48. Reprinted in *Alfred Stieglitz. Camera Work: The complete illustrations 1903–1917*, ed. Simone Philippi and Ute Kyeseyer, with an introduction by Pam Roberts, 106–107. Cologne: Taschen.
Stieglitz, A. 1897. The hand camera – its present importance. *American Annual of Photography and Photographic Times Almanac for 1897*: 18–27.

Cultural Information: Don't Ask, Don't Tell

Tim Lewens

1 Information in Cultural Evolution

Philosophers of biology have had plenty to say over the past dozen years or so about the notion of genetic information, especially since John Maynard Smith's (2000) agenda-setting paper. They have noted the widespread use of various apparently semantic notions within molecular and evolutionary biology: triplets *code* for amino acids; nucleotides constitute an *alphabet*; the genome contains *information* regarding proper development. Unsurprisingly, various different options have been put on the table for how to understand this kind of talk: genes embody naturalised information (e.g. Shea 2007 *inter alia*); talk of genetic information exemplifies a preformationist fallacy (Oyama 1985); the language of genetic information is a pernicious metaphor (Griffiths 2001); or an often useful fiction (Levy 2011).

Cultural evolutionists, too, speak freely of information. Indeed, they often tell us (with qualifications) that culture *is* information, as the following examples show. In his recent overview of cultural evolutionary theory, Alex Mesoudi says that "culture is information that is acquired from other individuals via social transmission mechanisms such as imitation, teaching or language" (2011, pp. 2–3). Here, Mesoudi is repeating a definition of culture that has been endorsed by a broad group of his collaborators – including Kevin Laland and Andrew Whiten – in a series of articles that have appeared over the past 10 years or so (e.g. Mesoudi et al. 2006, p. 331). This group, in turn, are following the lead of Boyd and Richerson, who define culture in a similar way as "information capable of affecting individuals'

T. Lewens (✉)
Department of History and Philosophy of Science, University of Cambridge,
Free School Lane, Cambridge CB2 3RH, UK
e-mail: tml1000@cam.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_25,

behaviour that they acquire from other members of their species through teaching, imitation, and other forms of social learning" (Richerson and Boyd 2005, p. 5).

We should not suppose that *only* cultural evolutionists have been attracted to informational conceptions of culture: for example, at the same time as expressing scepticism regarding the evolutionary frameworks of Richerson and Boyd (2005), Sperber (1996) and others, the cognitive anthropologist Maurice Bloch tells us that "What has been called culture is . . . a non-genetic, very long-term flow of information, in continual transformation, made possible by the fact that human beings are different from other animals because they can communicate to each other vast quantities of data, some of which they then may pass on to others" (Bloch 2012, p. 20). The appeal of the informational conception of culture extends beyond the boundaries of cultural evolutionary theory.

How are we to assess these informational visions of culture? In this paper, I argue for a "Don't Ask, Don't Tell" approach to cultural information. When we *ask* cultural evolutionists what they mean by "information", the answers they give us are problematic, and often these problems are very obvious. One might think, then, that the solution is for philosophers to *tell* cultural evolutionists what they ought to mean by cultural information, and more specifically that this should proceed via the adaptation of our best current theory of genetic information. Instead, I argue that the notion of cultural information as used by cultural evolutionists is not easily assimilated to teleofunctional notions of information that have been defended by the likes of Shea and others. Instead, we should aim for a scantily theorised notion of cultural information, which acts as an open-ended heuristic prompt, encouraging the cultural evolutionist to examine the ways in which bodies of skill, knowledge, norms and other cultural traits are reproduced from one generation to the next (Lewens 2014).

2 Cultural Information: Three Inadequate Accounts

In this Section I briefly examine, and criticize, three definitions of information offered within the cultural evolution community.

2.1 Richerson and Boyd (2005)

Richerson and Boyd *appear* to offer a definition of information in their *Not by Genes Alone*: "By information we mean any kind of mental state, conscious or not, that is acquired or modified by social learning and affects behavior" (2005, p. 5). This *looks* like a stipulative definition, but this cannot be the right way to interpret what they are up to. Elsewhere they stress that "some cultural information is stored in artifacts" (*ibid.*, p. 61). So whatever they mean by "information", they don't really mean to equate it with "mental state". In linking information to mental

states, they are merely expressing their empirical view that "cultural is (mostly) information stored in human brains" (*ibid.*, p. 61). They also take the view that culture may partially consist in information stored outside of human brains: they give the example of traditional pots. So Richerson and Boyd are not offering an account of what information is, just a claim about where most of it is.

The question of how we should understand "information" now becomes open again, as does the question of what culture – when conceived as information – is supposed to be. Richerson and Boyd's concession that some small proportion of cultural information might be stored in traditional pots is tantalizing, because it admits of (at least) two very different interpretations. They might be suggesting that these pots feature symbolic representations or inscriptions, i.e., that they store information in the everyday sense that a book stores information. But there is an alternative reading: the resources required for the transmission of pot-making are partly found in pots themselves. If old pots are used as models, and new pots are copied from them, then (whether pots have anything like intentionally encoded inscriptions or symbols on them) craftsmen rely on the structure of these old pots – and not just on their own portable know-how – when they come to make a new one. Hence some of the information regarding how to make a pot is to be found in pots.

This second reading is invited both by the nature of the example (it would be odd to mention pots, rather than books, if the point one wishes to get across is that some artefacts feature intentionally-coded information), and also by the context of discussion, for Boyd and Richerson are addressing the question of whether information stored in people's heads suffices to explain the reproduction of significant practices and objects. But once we agree that some information is stored in pots by virtue of their contribution to cultural reproduction, how far does cultural information extend? Suppose, as seems likely, that people learn how to milk cows by interacting with cows, perhaps through guided practice in the presence of an adept. An udder is required for the skill to be re-generated in later generations: just as the production of a new pot cannot take place without pots, so the production of the ability to milk a cow cannot take place without cows. Should we say that some cultural information is stored in cows? In which case, might one reasonably regard many other elements of the physical and biotic environments, which have also been affected by generations of humans, and which affect the development of future generations, as repositories of cultural information, too? Or is there some important asymmetry between the cow case and the pot case? Without further clarification of the nature of information, the extension of the evolutionary culture concept is unclear.

2.2 Boyd and Richerson (1985)

In their earlier work, Boyd and Richerson offer a more formal definition of information as "something which has the property that energetically minor causes have energetically major effects" (1985, p. 35). This definition is intended as a wholly

generic account of information: presumably it is meant to evoke intuitive examples whereby small informational "switches" (whether they are literally switches in a designed control system, or metaphorical "genetic switches" in developmental pathways) have magnified downstream effects on the systems they influence.

Boyd and Richerson take this account of information from an article by Engelberg and Boyarsky (1979), who explicitly have cybernetic control systems in mind when they write that "informational networks are characterized by (1) mapping, and (2) low-energy causes giving rise to high-energy effects" (p. 318). They form this view by reflecting on the fact that "the governor of a steam engine can be said to 'instruct' a valve to let in more or less steam", and they generalize by concluding that "causal links which we intuitively consider to be informational all have this amplificatory characteristic". However, there are plenty of cases of information-bearing relations where the energetic inequality is reversed. An instrument's display screen can carry information about solar flares: here, an energetically major cause has an energetically minor effect. As a general account of information, we can see that Engelberg and Boyarsky's picture won't do. It also won't do as a specific account of cultural information. In illustration of their concession that some small proportion of cultural information may reside outside people's heads, Richerson and Boyd (2005, p. 61) claim that church architecture contains information regarding the rituals one should perform. Their own definition of information would suggest that this can only be the case if it also turns out that a low energy cause here gives rise to a high energy effect. Passing briefly over the difficulties one might have in securing the claim that church architecture is a low energy cause compared with the high energy effect of performing a ritual, it seems that this account saddles the informational culture concept with an irrelevant epistemic hurdle. Surely it is not necessary to demonstrate this form of energetic inequality prior to claiming that some structure contains cultural information?

2.3 Hodgson and Knudsen (2010)

Hodgson and Knudsen's primary goal in their (2010) *Darwin's Conjecture* is to produce a generalised evolutionary theory suitable to all domains, from organic to cultural evolution. Their own particular area of interest is in developing an evolutionary economics, so to a degree they are peripheral to this paper's discussion of cultural evolution, but they have made more formal efforts to understand information and its role in evolutionary processes. Unlike Boyd and Richerson, they take the view that evolution – including cultural evolution – requires the existence of replicators, and they understand these replicators as bearers of information. What is useful for our purposes is their effort to specify an account of information that can discharge this task. We can begin with what I take to be the informal notion that lies behind their mathematized view. Intuitively, their hope is that we can assess the informational content – or "complexity" as they sometimes put it – of a replicator by comparing its own state with the state of a hypothetical

replicator that would maximise the fitness of the interactor that houses it in the actual replicator's environment. The actual environment of an actual replicator, in other words, specifies an ideal template for an optimal replicator, and any actual replicator is rich in information to the degree that it conforms with that optimal replicator.

There are many worries about this proposal: here, I focus on its informal aspect. Hodgson and Knudsen's approach is impractical, for it demands that we assess the information content of an ideal genome (or a body of know-how) that may not exist in any individual, prior to assessing whether real genomes (or bodies of know-how) are in alignment with this ideal. Even if such problems of measurement could be solved, there are more theoretical worries. Why should we think that there is any single optimal specification for one of these replicators? Might there not be several distinct replicator specifications that maximise interactor fitness? If so, how are we to assess distance between some actual replicator and the ideal?

This sort of problem becomes especially acute in the case of cultural evolution, for here the instability and manipulability of the environment of adaptation itself are especially vivid. A population in any given setting may have the capacity not merely to adapt to meet the demands of the environment, but to alter that environment so that it presents altered demands (Lewontin 1983; Odling-Smee et al. 2003; Lewens 2004, ch. 4). If we require some "ideal" specification for a cultural replicator before we can determine the informational content of an actual replicator, it seems we need some principled set of constraints on how that ideal replicator can be permitted to alter the environment in question. But Hodgson and Knudsen suggest no way of doing this, and I cannot see any way of achieving it.

3 The Perils of Information

In Sect. 2 we saw the deficiencies of definitions of information given by cultural evolutionary theorists. Here, I show that confusions over the nature of cultural information cause a variety of problems in cultural evolutionary theorising. Mesoudi, for example, never gives an explicit definition of information, but he does give a list of examples. Information, he says, is "intended as a broad term to refer to what social scientists and lay people might call knowledge, beliefs, attitudes, norms, preferences, and skills, all of which may be acquired from other individuals via social transmission and consequently shared across social groups" (2011, p. 3). While Mesoudi is quite liberal in the sorts of things that can count as forms of information, he nonetheless stresses that culture is "information rather than behaviour" (*ibid.*). Grant Ramsey, in a very recent paper, has also denied that culture should be thought of as behaviour, on the grounds that "if culture is behavior, then culture cannot cause or explain behavior" (Ramsey 2013). Mesoudi's reasons for excluding behaviour from the definition of culture are similar to Ramsey's – he is concerned about avoiding circular forms of explanation – but this form of argument fails, because it proves too much (Lewens 2012).

As we have seen, Mesoudi takes it that "information" names a variety of states that include "knowledge" and "skill", so Mesoudi's version of the circularity objection would appear to have the consequence that culture cannot explain knowledge and skill; and yet, cultural evolutionists typically require that culture can explain such things. Ramsey recognizes the logical strength of the circularity argument, and he consequently bites the bullet, requiring not merely that culture should not be defined in terms of behaviour, but that it should not be defined in terms of any cognitive states, including belief and knowledge. He therefore requires some notion of culture as information that puts conceptual distance between informational states and cognitive states. But there is really no problem of circularity: even if "culture" names a variety of cognitive or behavioural states, culture can still explain such states on the grounds that the cognitive or behavioural endowment of one generation (i.e., its culture) can contribute to the production of a resembling cognitive or behavioural endowment in another generation.

Mesoudi adds to these worries about circularity the thought that "there are other causes of behaviour besides culture" (2011, p. 4). For example, behaviours may be caused genetically or by individual learning. If, then, our definition of culture is to make room for alternative causes of behaviour, Mesoudi concludes that we had better not make behaviour a subspecies of culture. Again, though, this objection proves too much: it would also have the consequence that since knowledge and skills – or, for that matter, information – can potentially be produced by processes such as individual learning, we had better not equate culture with knowledge, skills, or even with information more generally. And yet, these are just the equations Mesoudi and Ramsey wish to make.

Further worries about how we are to understand the notion of information are exemplified by the tension we have already seen between Richerson and Boyd's initial claim that a necessary condition for "cultural information" is that it names a "kind of mental state" (albeit one acquired in a certain kind of way, via some form of social learning), and their follow-up claim that some "cultural information" can reside in pots. Perhaps this confusion occurs because of the conceptual baggage of notions of genetic inheritance; Richerson and Boyd argue that cultural variants "must be genelike to the extent that they carry cultural information" (2005, p. 81). Genetic information is sometimes thought to direct organic reproduction. Perhaps trading on this aspect of the genetic domain, the term "cultural information" now serves double duty: on the one hand, it inherits from this notion of genetic information the role of directing cultural reproduction; on the other hand, "cultural information" names a loose subset of mental states, i.e., the mental states which are themselves reproduced by "cultural information" as understood in the first sense. But of course, having isolated a notion of culture as a subset of mental states, the question of what explains the reliable reproduction of those states is left open: perhaps what explains the reliable reproduction of pot-making skills in part lies in the material constitution of pots; perhaps what explains the reliable reproduction of distinctive food preferences lies partly in the stable availability of a restricted set of foods in a community's locale. It is not surprising that tensions emerge in a culture concept that is supposed to work both as explanandum and as explanans.

4 Looking to Philosophy

So far, I have considered the efforts of practitioners within the domain of cultural evolution to define the informational culture concept. In a moment, I will turn to a potentially more promising source for a rigorous account of information, by looking at the efforts of philosophers to construct an information concept suitable to informational understandings of genetic inheritance. Before I do this, I want to flag some reasons that might make us sceptical about how likely it is that philosophical accounts fashioned primarily in order to make sense of genetic information are likely to be up to the job of elucidating cultural information.

First, the historical contexts in which information talk gets in to genetic and cultural evolution are very different. Knusden (not to be confused with Knudsen, whom we met in Sect. 2) has argued that in the early years of research in molecular biology, during the 1940s and 1950s, semantic metaphors played an important heuristic role in formulating plausible hypotheses for the relationship between chromosomal structure and developmental pathways (Knusden 2005). Schrödinger, for example, wrote that

> In calling the structure of the chromosome fibres a code-script we mean that the all-penetrating mind, once conceived by Laplace, to which every causal connection lay open, could tell from their structure whether the egg would develop, under suitable conditions, into a black cock or into a speckled hen, into a fly or a maize plant, a rhododendron, a beetle, a mouse or a woman (Schrödinger 1944, p. 21).

Schrödinger did not have access to the details of how proteins are produced; instead, he gestures towards the hypothesis that the chromosomes contain some kind of code, which in turn is a way of saying that their configuration alone would allow one to predict how development will proceed. Rahul Rose (2012) has pointed out that this type of talk was often explicitly metaphorical, at the same time as it pointed to some possible interpretation in more neutral biochemical language. So (borrowing Rose's own examples), Goldstein and Plaut hedged their use of semantic language within scare quotes: "It has been suggested that RNA could serve as a receptor of a 'code' from DNA in the nucleus and could transmit this specificity to cytoplasmic proteins, with the synthesis of which it may be associated" (1955, p. 874). Levinthal stressed the conjectural nature of the code hypothesis when he claimed that "although there is some evidence that a code exists which translates genetic information into amino-acid sequence, there is no information as to the *nature* of this code" (1959, p. 254, emphasis in original).

Informational language initially found its way into genetics via a bold and contentious hypothesis: the chromosomal material somehow embodies a code that specifies amino-acid sequence, or perhaps protein structure. The problem, then, was to find out whether there was indeed such a code, and what that code was. Clearly, no such genealogy can be found when we turn to the notion of cultural information. No one would think it a bold conjecture to be told that cultural practices are produced by the action of some form of cultural code. In some areas the suggestion is too obviously true, in others it seems obviously false. On the one hand, we think it

trivial that cultures do contain written codes: it would hardly be news if a cultural evolutionist told us that languages might have a code-like structure. On the other hand, even if we acknowledge the obvious point that we learn from observing others, it is very hard indeed to see how one might think of, say, elements of a demonstration of flint-knapping mapping onto neural patterns in the observer, or onto final acquired skills, in a code-like manner in the same way that one thinks of the ordering of a nucleotide triplet mapping on to the amino acid to be produced. The results of instruction do not seem stable enough for any kind of coding hypothesis to be attractive. It is hardly surprising, then, that while the language of molecular biology is saturated with the language not just of code, but of proof-reading, transcription, translation, and so forth, no such detailed semantic framework has been incorporated into theorising about cultural evolution.

The heritage of the notion of cultural information is quite different to the heritage of the notion of genetic information. Cultural evolutionary theorists talk of cultural information for three main reasons. First, cultural evolutionists focus in large part on changes in human mental states over time. Mental states are paradigmatic information-bearing states: it is hardly problematic to think that knowledge, for example, involves the possession of information. Second, cultural evolutionary theorists note that one of the things our species is adept at is acquiring knowledge from others through learning processes: so talk of the transmission of information is entirely natural. Third, cultural evolutionists have a significant interest in understanding the processes by which knowledge of various forms is retained and changed: it is not surprising that this becomes framed, informally, as a project to uncover the flows of information that allow the sustenance of cognitive capital (e.g. Sterelny 2012). The cultural evolutionist's main business is to understand how a collection of intentional states comes to change over time under the influence of forms of social transmission. That is why, in lieu of a definition, we now understand why Mesoudi instead says, entirely reasonably, that Information is "intended as a broad term to refer to what social scientists and lay people might call knowledge, beliefs, attitudes, norms, preferences, and skills, all of which may be acquired from other individuals via social transmission and consequently shared across social groups" (2011, p. 3). Much of Richerson and Boyd's work makes a similar compromise: they frequently use the term "cultural variant" as a shorthand for "information stored in people's heads" (2005, p. 63). But they also offer the alternative of "the ordinary English words idea, skill, belief, attitude, and value" (*ibid.*). In the great majority of their work, they simply seek to model the ways in which different learning biases can affect the populational profile of a collection of ideas, skills, beliefs, attitudes and so forth. "Cultural information", when used in this context, is nothing more than a shorthand for this broad collection of phenomena. This project requires no formal specification of what cultural information is supposed to be and, as we have seen, when theorists are tempted to offer such a specification it typically lands them in trouble.

5 Don't Tell

I have just argued that "cultural information" is often simply used as a shorthand to denote a variety of mental states, including beliefs, skills, preferences and so forth. Sometimes "cultural information" is instead used in a rather different way. Recall that Richerson and Boyd tell us that cultural information need not be stored wholly in people's heads:

> Undoubtedly some cultural information is stored in artifacts. The designs that are used to decorate pots are stored on the pots themselves, and when young potters learn how to make pots they use old pots, not old potters, as models. In the same way the architecture of the church may help store information about the rituals performed within (2005, p. 61).

They are telling us two things here: first, there is a sense of "cultural information" that allows that some of it, at least, can be stored in pots, architectural structures, and so forth. They are also suggesting that this counts as information because it helps to explain the reproduction of cultural practices. The problem is that they aren't telling us much more than this and one might worry that, without further specification, the potential repositories of cultural information are entirely open-ended.

It is at this point that one might think it useful to look to philosophical theories to tell cultural evolutionists what they ought to mean by the notion of "cultural information". There are many options one could turn to, but here I will focus on Nick Shea's "infotel" theory (e.g. Shea 2007, 2013). Shea argues that genes can be said to contain information because it is the evolved function of the genome to bring about heritable variation. Shea's theory builds on, and complexifies, earlier teleosemantic accounts of genetic information due to Maynard-Smith (2000) and Sterelny et al. (1996). His theory allows us to say, in a full-blooded manner, that genes represent environmental conditions, that they can be misread, that they contain instructions and so forth. Shea's theory allows that many developmental resources can interact to explain development and reproduction, while only some of those resources contain information about developmental outcomes. To give just one example, a suitable gravitational field is required for offspring to resemble parents, but there has been no selective history whereby the properties of the gravitational field have been altered in such a way that improves its ability to stabilize parent/offspring resemblance. On the other hand, the existence of forms of proof-reading machinery, and so forth, do provide strong evidence that the functional role of the genome is to bring about these stable resemblances.

Shea's overall theory has significant heuristic payoffs: it is undeniably useful to ask, for any process whereby offspring come to resemble parents, or more generally where one generation comes to resemble another, whether an "inheritance system" is at work, and whether it has the sorts of features that appear to indicate complex adaptation as an inheritance system. This kind of framework helps us to mobilise a series of important general reflections regarding the costs and benefits of general features of inheritance for adaptive evolution. That said, even as we acknowledge the

heuristic benefits of asking after the functional constraints on inheritance systems, we should deny that Shea puts his finger on the way in which "information" is used in cultural evolutionary theory.

A couple of brief examples from cultural evolutionary theorizing suggest it may be unwise to apply Shea's theory in this domain. Kim Sterelny's very recent work argues that skills in an offspring generation may sometimes be acquired because parents (i) engage in skilled activities, (ii) their engagement in these activities results in various tools, raw materials and so forth "lying around" for experimentation by others, and (iii) offspring therefore get to experiment with these pre-prepared objects in ways that make their own acquisition of the relevant skills easier (Sterelny 2012). This is just the sort of set-up that one may wish to describe – and that Sterelny does describe – in terms of the flow of information from parental generation to offspring generation. The problem, however, is that Shea's appeal to evolved functions in determining what counts as an inheritance system, and his definition of informational states in terms of inheritance systems, together have the result that *prior* to the advent of specializing adaptations that improve inheritance, it is strictly inappropriate to speak of cultural information at all. And yet, Sterelny wishes to stress how valuable transmission of cultural information can occur without such specialization, and he wishes to point out that these forms of highly disaggregated information transfer may have been important in the early stages of our cognitive development. Sterelny's story for human evolutionary adaptation is precisely one in which we become better and better adapted at making use of cultural information. The sort of story he has in mind begins with "accidental" information transmission, whereby juveniles simply hang around with their parents, and the result is that they take advantage of an environment that is structured in such a way that they have both learning opportunities, and suitable materials with which to learn. This structure has been produced by the social action of a previous generation. Subsequent adaptive steps may include greater tolerance of experimenting juveniles, and eventually something like explicit apprenticeship. So Sterelny is thoroughly attuned to the ways in which the design features of cultural inheritance systems may improve over time.

We can salvage something from Shea's account, which stresses the kinship between his view and metaphor theories of genetic and cultural information (e.g. Levy 2011). Shea has a nice way of expressing the motivating intuition behind his theory of information (Shea 2012). The bills of purple-throated carib hummingbirds develop to match the flowers they feed from. This "good match" does not derive from causal interactions with the flowers in question during development: "How does the developing hummingbird 'know' what shape of bill to produce, to match available nectar sources in its local ecology?" (2012, p. 2237). Shea's idea is that on some occasions, this "knowledge" or "information" is contained in the hummingbird genome, and generated through cycles of selection across generations.

Informational talk is indeed encouraged when one asks, metaphorically speaking, "how does this organism know how to behave?", or "where does the information come from?" In a widely cited paper by Danchin et al. (2004) on the impact of "public information" on cultural evolution, the authors make use of a framework in which various forms of "non-genetic information" are discussed. The only definition they give to "information" more generally is extremely cursory: it is defined as

"anything that reduces uncertainty" (*id*., p. 487). Of course, this alludes to a well-known account of Shannon information. But Danchin et al., do not literally mean to suggest that the uptake of information reduces the cognitive uncertainty of the animals they discuss. Information helps resolve uncertainty in the organism, in the sense that it causes some outcome that is appropriate in the circumstances. For that reason, one might just as well talk of a plant's uncertainty regarding how to develop being reduced by the presence of an environmental correlate of impending drought.

Importantly, Danchin et al., distinguish *signals* – "traits specifically designed by selection to convey information" (2004, p. 487) – from "Inadvertent Social Information". In these latter cases, the behaviour of organisms leaves a trail of relevant clues, which can be used by conspecifics even though such behaviours are *not* modified by selection for these functions of transmission. Danchin et al., give several examples, including the ways in which choice of habitat or mating site might be affected by a conspecific interacting with the behaviour of others, or interacting with the downstream effects of another's behaviour. These clues help to inform the organism regarding how to act, or how to develop. In some cases, then, behaviours are described as carrying information, in spite of the fact that these behaviours do not have the teleofunction of carrying information.

Any factor that "reduces uncertainty" might be considered a source of "information" feeding in to the broad process of appropriate development. There are a number of potential information sources: suitable developmental responses may come from innate representation, from explicit instruction, from imitation of another. Another potential explanation, suggested in Sterelny's discussion of the acquisition of skills by juveniles, adverts to interaction with an environment that is already well structured for the generation of the skill in question. In all cases, talk of information becomes attractive whenever development, or behaviour, can be interpreted in such a way that it is guided in a suitable manner by the presence or absence of material factors, where these might include the configuration of genomic elements, the placement of raw materials for construction of an axe, the mating behaviours of conspecifics, and so forth. Shea's metaphorical question, "how does the organism know what to do?", needn't be answered by appeal to a structure that has been shaped by selection in order to facilitate resemblance with respect to adaptive characters, even though sometimes it will be answered in this way. I suggest, then, that searches for sources of cultural information are typically unencumbered by requirements that the structures bearing information have teleofunctions of inheritance (see also Levy 2011).

The account offered here is therefore at odds with Shea's view, and only partially compatible with some widely discussed comments on information by Bergstrom and Rosvall (2011). Bergstrom and Rosvall focus on a notion of information, according to which information is something transmitted from one generation to another. They appear to require that information-bearing structures have naturalized teleofunctions of transmission. As they put it, "Like naturalized views of semantics, the transmission notion of information rests upon function: to say that X carries information, we require that the function of X be to reduce uncertainty on the part of a receiver" (*id*., p. 169). A little earlier in the article, the teleofunctional requirement

is even more explicit: "Our aim with the transmission sense of information is [. . .] to identify those components of biological systems that have information storage, transmission, and retrieval as their primary function" (*id*., p. 167). They ask:

> So must we impose our own notions of what makes an appropriate reference frame in order to single out certain components of the developmental matrix as signal and others as noise? If we want to know how the information necessary for life was compiled by natural selection, the answer is no. In this case, we are not the ones who pick the reference frame, natural selection is (*id*., p. 168).

This passage begs the important question of whether the information necessary for life is indeed always compiled by natural selection: why could it not be the case that some information necessary for life is compiled in some other way, perhaps by the intentional structuring of a learning environment, but perhaps as the accidental by-product of artisanal activities, or mating activities? Earlier in their article, Bergstrom and Rosvall remind us of a more general account of information: "An object X conveys information if the function of X is to reduce, by virtue of its sequence properties, uncertainty on the part of an agent who observes X" (*id*., p. 165). If we make use of this definition, we can dispense with the teleofunctional condition on information, and instead construe the function of an object contributing to development via a covert assumption that we should regard the organism *as if* it were an observing agent making use of the various resources to hand as it develops. Anything used by the organism for this purpose would then acquire the function of reducing the organism's uncertainty. There is even a suggestion that Bergstrom and Rosvall have something like this in mind themselves: consider their comment that "A single individual can only look at its own genome and see a sequence of base pairs. This sequences of base pairs is what is transmitted; it is what has the function of reducing uncertainty on the part of the agent who observes it" (*id*., pp. 169–170). Of course, an organism does not literally look at its own genome, but its genome, just like various aspects of its social and technical environments, can affect its developmental trajectory in an adaptive manner. It will then be entirely legitimate to regard Danchin et al.'s "Inadvertent Social Information", or Sterelny's structured learning environments, as bona-fide loci of information too, in spite of the fact that their own structurings have not been selected for the function of transmission. Moreover, we can then move on to ask, of any of these informational loci, the sorts of important heuristic questions Bergstrom and Rosvall recommend to us regarding channel capacities, redundancy, and so forth. We can also ask the sorts of questions recommended by Maynard-Smith and Szathmary (1995) regarding the comparative benefits of different informational channels. The vital heuristic functions of Bergstrom and Rosvall's "transmission sense" of information do not require a teleofunctional underpinning.

6 Two Faces of Cultural Information

The notion of information suggested by teleosemantics is surplus to the requirements of cultural evolutionary theory. For the most part, cultural evolutionists can use "cultural information" as a shorthand for their main explanatory target, which is

distributions of mental states. Sometimes, cultural evolutionists are also prompted to ask how these distributions are sustained and altered. Here, the search for cultural information is the search for those factors that explain the organism's dispositions to behave appropriately: loosely, it is the search for factors that help to answer Shea's question, 'how does the organisms know what to do?' This might involve an appeal to formal teaching, or to individual learning, or to the structuring of learning environments, or to the laying down of written documents or to the development of innate skills.

Not just any informational resources are labelled as *cultural*: a suitable allocation of genes is required for the re-development of cultural practices, but one would not typically say that genes contain cultural information. "Cultural information", then, labels the *cultural* resources that enable the reproduction of cultural states. Can we say which are the "cultural resources", as opposed to the non-cultural ones? Here is a suggestion: informational resources tend to be labeled as "cultural" when they affect behaviour via learning (hence marks on pots count, but marks on chromatin do not), and when their own distribution is under the influence of social and cognitive processes (hence wild medicinal plants do not contain cultural information, but artificial ecclesiastical structures do). How, now, does this handle the problems of cows, mentioned back in Sect. 2? Does this mean we must say that cow's udders contain cultural information, on the grounds that a pastoralist learns how to milk a cow by exposure to cows' udders, and on the grounds that the makeup and distribution of cows' udders is under social control? Perhaps this does sound strange, but this type of awkward consequence is hardly damaging to a broad theory that seeks to document the ways in which socially structured environments interact with learning dispositions to enable the transmission of valuable skills and beliefs from one generation to the next.

Acknowledgements An earlier version of this paper was given at the conference "Philosophy and the Sciences: Old Visions, New Directions" in Cambridge in December 2012. I am grateful to the audience there, and especially to Paul Griffiths, for comments. For further discussion I am grateful to Beth Hannon and Andrew Buskell. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no 284123.

References

Bergstrom, C., and M. Rosvall. 2011. The transmission sense of information. *Biology and Philosophy* 26: 159–176.

Bloch, M. 2012. *Anthropology and the cognitive challenge*. Cambridge: Cambridge University Press.

Boyd, R., and P.J. Richerson. 1985. *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Danchin, E., L. Giraldeau, T. Valone, and R. Wagner. 2004. Public information: From nosy neighbours to cultural evolution. *Science* 305: 487–491.

Engelberg, J., and L. Boyarsky. 1979. The noncybernetic nature of ecosystems. *American Naturalist* 114: 327–324.
Goldstein, L., and W. Plaut. 1955. Direct evidence for nuclear synthesis of cytoplasmic ribose nucleic acid. *PNAS* 41: 874–880.
Griffiths, P. 2001. Genetic information: A metaphor in search of a theory. *Philosophy of Science* 68: 394–412.
Hodgson, J., and T. Knudsen. 2010. *Darwin's conjecture: The search for general principles of social and economic evolution*. Chicago: University of Chicago Press.
Knusden, S. 2005. Communicating novel and conventional scientific metaphors: A study of the development of the metaphor of genetic code. *Public Understanding of Science* 14: 373–392.
Levinthal, C. 1959. Coding aspects of protein synthesis. *Reviews of Modern Physics* 31: 249–255.
Levy, A. 2011. Information in biology: A fictionalist account. *Nous* 45: 640–657.
Lewens, T. 2004. *Organisms and artifacts: Design in nature and elsewhere*. Cambridge, MA: MIT Press.
Lewens, T. 2012. The Darwinian view of culture. *Biology and Philosophy* 27: 745–753.
Lewens, T. 2014. Review of the evolved apprentice. *British Journal for the Philosophy of Science* 65: 185–189.
Lewontin, R. 1983. The organism as the subject and object of evolution. *Scientia* 118: 63–82.
Maynard Smith, J. 2000. The concept of information in biology. *Philosophy of Science* 67: 177–194.
Maynard Smith, J., and E. Szathmary. 1995. *The major transitions in evolution*. Oxford: Oxford University Press.
Mesoudi, A. 2011. *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. Chicago: University of Chicago Press.
Mesoudi, A., A. Whiten, and K.N. Laland. 2006. Towards a unified science of cultural evolution. *Behavioral and Brain Sciences* 29: 329–347.
Odling-Smee, J., K.N. Laland, and M. Feldman. 2003. *Niche construction: The neglected process in evolution*. Princeton: Princeton University Press.
Oyama, S. 1985. *The ontogeny of information: Developmental systems and evolution*. Cambridge, MA: Cambridge University Press.
Ramsey, G. 2013. Culture in humans and other animals. *Biology and Philosophy* 28: 457–479.
Richerson, P.J., and R. Boyd. 2005. *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.
Rose, R. 2012. *Semantic information in genetics: A metaphorical reading*. Unpublished M.Phil. essay, University of Cambridge, Department of History and Philosophy of Science.
Schrödinger, E. 1944. *What is life?* Cambridge: Cambridge University Press.
Shea, N. 2007. Representation in the genome and in other inheritance systems. *Biology and Philosophy* 22: 313–331.
Shea, N. 2012. New thinking, innateness and inherited representation. *Philosophical Transactions of the Royal Society B* 367: 2234–2244.
Shea, N. 2013. Inherited representations are read in development. *British Journal for the Philosophy of Science* 64: 1–31.
Sperber, D. 1996. *Explaining culture: A naturalistic approach*. Oxford: Blackwell.
Sterelny, K. 2012. *The evolved apprentice*. Cambridge, MA: MIT Press.
Sterelny, K., K. Smith, and M. Dickison. 1996. The extended replicator. *Biology and Philosophy* 11: 377–403.

Part IV
Philosophy of the Physical Sciences

Introducing QBism

Christopher A. Fuchs

1 A Feared Disease

The end of the last decade saw a media frenzy over possibility of an H1N1 flu pandemic. The frenzy turned out to be misplaced, but it did serve to remind us of a basic truth: that a healthy body can be stricken with a fatal disease which to outward appearances is nearly identical to a common yearly annoyance. There are lessons here for quantum mechanics. In the history of physics, there has never been a healthier body than quantum theory; no theory has ever been more all-encompassing or more powerful. Its calculations are relevant at every scale of physical experience, from subnuclear particles, to table-top lasers, to the cores of neutron stars and even the first 3 min of the universe. Yet since its founding days, many physicists have feared that quantum theory's common annoyance – the continuing feeling that something at the bottom of it does not make sense – may one day turn out to be the symptom of something fatal.

There is something about quantum theory that is different in character from any physical theory before. To put a finger on it, the issue is this: the basic statement of the theory – the one we have all learned from our textbooks – seems to rely on terms our intuitions balk at as having any place in a fundamental description of reality. The notions of "observer" and "measurement" are taken as primitive, the very starting point of the theory. This is an unsettling situation! Shouldn't physics

C.A. Fuchs (✉)
Perimeter Institute for Theoretical Physics, 31 Caroline Street North,
N2L 2Y5 Waterloo, Ontario, Canada
e-mail: cfuchs@perimeterinstitute.ca

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_26,

be talking about *what is* before it starts talking about *what will be seen* and who will see it? Few have expressed this more forcefully than John Stewart Bell (Bell 1990):

> What exactly qualifies some physical systems to play the role of 'measurer'? Was the wavefunction of the world waiting to jump for thousands of millions of years until a single-celled living creature appeared? Or did it have to wait a little longer, for some better qualified system ... with a PhD?

One sometimes feels that until this issue is settled, fundamental physical theory has no right to move on.

But what constitutes "progress" in quantum foundations? How would one know progress if one saw it? Through the years, it seems the most popular strategy has taken its cue (even if only subliminally) from the tenor of John Bell's quote: the idea has been to remove the observer from the theory just as quickly as possible, and with surgical precision. In practice this has generally meant to keep the *mathematical structure* of quantum theory as it stands (complex Hilbert spaces, operators, tensor products, etc.), but, by hook or crook, invent a story about the mathematical symbols that involves no observers at all.

In short, the strategy has been to reify or objectify all the mathematical symbols of the theory and then explore whatever comes of the move. Three examples suffice to give a feel: in the de Broglie – Bohm "pilot wave" version of quantum theory, there are no fundamental measurements, only "particles" flying around in a $3N$-dimensional configuration space, pushed around by a wave function regarded as a real physical field in that space. In "spontaneous collapse" versions, systems are endowed with quantum states that generally evolve unitarily, but from occasionally collapse without any need for measurement. In Everettian or "many-worlds" quantum mechanics, it is only the world as a whole – they call it a multiverse – that is really endowed with an intrinsic quantum state, and that quantum state evolves deterministically, with only an *illusion from the inside* of probabilistic "branching".

The trouble with all these interpretations as quick fixes for Bell's stirring remark is that they look to be just that, *really quick fixes*. They look to be interpretive strategies hardly compelled by the particular details of the quantum formalism, giving only more or less arbitrary appendages to it. This already explains in part why we have been able to exhibit three such different strategies, but it is worse: each of these strategies gives rise to its own set of incredibilities – ones which, if one had Bell's gift for words, one could make look just as silly. Pilot-wave theories, for instance, give instantaneous action at a distance, but not actions that can be harnessed to send detectable signals. If so, then what a delicately balanced high-wire act nature presents us with. Or the Everettians. Their world purports to have no observers, but then it has no probabilities either. What are we then to do with the Born Rule for calculating quantum probabilities? Throw it away and say it never mattered?

2 Quantum States Do Not Exist

There was another lesson from the H1N1 virus. It is that sometimes immunities can be found in unexpected populations. To some perplexity at the beginning, it seemed that people over 65 – a population usually more susceptible to fatalities with seasonal flu – fared much better than younger folk with H1N1. The leading theory was that the older population, in its years of other exposures, has developed various latent antibodies. The antibodies were not perfect, but they were a start. And so it may be for quantum foundations.

Here, the latent antibody is the concept of *information*, and the perfected vaccine, we believe, will arise in part from the theory of single-case, personal probabilities – the branch of probability theory called Bayesianism. Symbolically, the older population corresponds to some of the very founders of quantum theory (Heisenberg, Pauli, Einstein) and some of the younger disciples of the Copenhagen school (Peierls, Wheeler, Peres), who, though they disagreed on many details of the vision – *Whose information? Information about what?* – were unified on one point: that quantum states are not something out there, in the external world, but instead are expressions of information. Before there were people using quantum *theory* as a branch of physics, before they were *calculating* neutron-capture cross-sections for uranium and working on all the other practical problems the theory suggests, there were no quantum states. The world may be full of stuff and things of all kinds, but among all the stuff and all the things, there is no unique, observer-independent, *quantum-state kind of stuff*.

The immediate payoff of this strategy is that it eliminates the conundrums arising in the various objectified-state interpretations. A paraphrase of a quote by James Hartle makes the point decisively (Hartle 1968):

> A quantum-mechanical state being a summary of the observers' information about an individual physical system changes both by dynamical laws, and whenever the observer acquires new information about the system through the process of measurement. The existence of two laws for the evolution of the state vector becomes problematical only if it is believed that the state vector is an objective property of the system. If, however, the state of a system is defined as a list of [experimental] propositions together with their [probabilities of occurrence], it is not surprising that after a measurement the state must be changed to be in accord with [any] new information. The 'reduction of the wave packet' does take place in the consciousness of the observer, not because of any unique physical process which takes place there, but only because the state is a construct of the observer and not an objective property of the physical system.

It says that the real substance of Bell's fear is just that, the fear itself. To succumb to it is to block the way to understanding the theory on its own terms. Moreover, the shriller notes of Bell's rhetoric are the least of the worries: the universe didn't have to wait billions of years to collapse its first wave function – wave functions are not part of the observer-independent world.

But this much of the solution is an elderly and somewhat ineffective antibody. Its presence is mostly a call for more clinical research. Luckily the days for this are ripe, and it has much to do with the development of the field of quantum information

theory – the multidisciplinary field that brought about quantum cryptography, quantum teleportation, and will one day bring about quantum computation. What the protocols and theorems of quantum information pound home is the idea that quantum states look, act, and feel like information in the technical sense of the word – the sense provided by probability theory and Shannon's information theory.

There is no more beautiful demonstration of this than Robert Spekkens's "toy model" for mimicking various features of quantum mechanics (Spekkens 2007). In that model, the "toys" are each equipped with four possible mechanical configurations; but the players, the manipulators of the toys, are consistently impeded – for whatever reason! – from having more than 1 bit of information about each toy's actual configuration. (Or a total of 2 bits for each 2 toys, 3 bits for each 3 toys, and so on.) The only things the players can know are their states of uncertainty about the configurations. The wonderful thing is that these states of uncertainty exhibit many of the characteristics of quantum information: from the no-cloning theorem to analogues of quantum teleportation, quantum key distribution, entanglement monogamy, and even interference in a Mach-Zehnder interferometer. More than two dozen quantum phenomena are reproduced *qualitatively*, and all the while one can always pinpoint the underlying cause of the occurrence: the phenomena arise in the uncertainties, never in the mechanical configurations. It is the states of uncertainty that mimic the formal apparatus of quantum theory, not the toys' so-called *ontic states* (states of reality).

What considerations like this tell the ψ-ontologists – i.e., those who to attempt to remove the observer too quickly from quantum mechanics by giving quantum states an unfounded ontic status – was well put by Spekkens:

> [A] proponent of the ontic view might argue that the phenomena in question are not mysterious if one abandons certain preconceived notions about physical reality. The challenge we offer to such a person is to present a few simple physical principles by the light of which all of these phenomena become conceptually intuitive (and not merely mathematical consequences of the formalism) within a framework wherein the quantum state is an ontic state. Our impression is that this challenge cannot be met. By contrast, a single information-theoretic principle, which imposes a constraint on the amount of knowledge one can have about any system, is sufficient to derive all of these phenomena in the context of a simple toy theory ...

The point is, far from being an appendage cheaply tacked on to the theory, the idea of quantum states as information has a simple unifying power that goes some way toward explaining why the theory has the very mathematical structure it does. By contrast, who could take the many-worlds idea and derive any of the structure of quantum theory out of it? This would be a bit like trying to regrow a lizard from the tip of its chopped-off tail: the Everettian conception never purported to be more than a reaction to the formalism in the first place.

There are, however, aspects of Bell's challenge (or the mindset behind it), that remain a worry. And upon these, all could still topple. There are the old questions of *Whose information?* and *Information about what?* – these certainly must be addressed before any vaccination can be declared a success. It must also be settled whether quantum theory is obligated to give a criterion for what counts as an

observer. Finally, because no one wants to give up on physics, we must tackle head-on the most crucial question of all: if quantum states are not part of the stuff of the world, then what is? What sort of stuff does quantum mechanics say the world *is* made of?

Good immunology does not come easily. But this much is sure: the glaringly obvious (that the central part of quantum theory is about information) should not be abandoned rashly: to do so is to lose grip of the theory as it is applied in practice, with no better grasp of reality in return. If on the other hand, one holds fast to the central point about information, initially frightening though it may be, one may still be able to reconstruct a picture of reality from the unfocused edge of vision. Often the best stories come from there anyway.

3 Quantum Bayesianism

Every area of human endeavor has its bold extremes. Ones that say, "If this is going to be done right, we must go this far. Nothing less will do." In probability theory, the bold extreme is the personalist Bayesian account of it (Bernardo and Smith 1994). It says that probability theory is of the character of formal logic – a set of criteria for testing consistency. In the case of formal logic, the consistency is between truth values of propositions. However logic itself does not have the power to *set* the truth values it manipulates. It can only say if various truth values are consistent or inconsistent; the actual values come from another source. Whenever logic reveals a set of truth values to be inconsistent, one must dip back into the source to find a way to alleviate the discord. But precisely in which way to alleviate it, logic gives no guidance. "Is the truth value for this one isolated proposition correct?" Logic itself is powerless to say.

The key idea of personalist Bayesian probability theory is that it too is a calculus of consistency (or "coherence" as the practitioners call it), but this time for one's decision-making degrees of belief. Probability theory can only say if various degrees of belief are consistent or inconsistent with each other. The actual beliefs come from another source, and there is nowhere to pin their responsibility but on the agent who holds them. As Dennis Lindley put it (Lindley 2006),

> The Bayesian, subjectivist, or coherent, paradigm is egocentric. It is a tale of one person contemplating the world and not wishing to be stupid (technically, incoherent). He realizes that to do this his statements of uncertainty must be probabilistic.

A probability *assignment* is a tool an agent uses to make gambles and decisions – it is a tool he uses for navigating life and responding to his environment. Probability *theory* as a whole, on the other hand, is not about a single isolated belief, but about a whole mesh of them. When a belief in the mesh is found to be incoherent with the others, the theory flags the inconsistency. However, it gives no guidance for how to mend any incoherences it finds. To alleviate the discord, one can only dip back into the source of the assignments – specifically, the agent who attempted to sum up all

his history, experience, and expectations with those assignments in the first place. This is the reason for the terminology that a probability is a "degree of belief" rather than a "degree of truth" or "degree of facticity."

Where personalist Bayesianism breaks away the most from other developments of probability theory is that it says there are no *external* criteria for declaring an isolated probability assignment right or wrong. The only basis for a judgment of adequacy comes from the *inside*, from the greater mesh of beliefs the agent may have the time or energy to access when appraising coherence.

It was not an arbitrary choice of words to title the previous section

QUANTUM STATES DO NOT EXIST,

but a hint of the direction we must take to develop a perfected vaccine. This is because the phrase has a precursor in a slogan Bruno de Finetti, the founder of personalist Bayesianism, used to vaccinate probability theory itself. In the preface of his seminal book de Finetti (1990), de Finetti writes, centered in the page and in all capital letters,

PROBABILITY DOES NOT EXIST.

It is a powerful statement, constructed to put a finger on the single most-significant cause of conceptual problems in pre-Bayesian probability theory. A probability is not a solid object, like a rock or a tree that the agent might bump into, but a feeling, an estimate inside himself.

Previous to Bayesianism, probability was often thought to be a physical property – something objective and having nothing to do with decision-making or agents at all. But when thought so, it could be thought only inconsistently so. And hell hath no fury like an inconsistency scorned. The trouble is always the same in all its varied and complicated forms: if probability is to be a physical property, it had better be a rather ghostly one – one that can be told of in campfire stories, but never quite prodded out of the shadows. Here's a sample dialogue:

Pre-Bayesian: Ridiculous, probabilities are without doubt objective. They can be seen in the relative frequencies they cause.

Bayesian: So if $p = 0.75$ for some event, after 1,000 trials we'll see exactly 750 such events?

Pre-Bayesian: You might, but most likely you won't see that exactly. You're just likely to see something close to it.

Bayesian: Likely? Close? How do you define or quantify these things without making reference to your degrees of belief for what will happen?

Pre-Bayesian: Well, in any case, in the infinite limit the correct frequency will definitely occur.

Bayesian: How would I know? Are you saying that in one billion trials I could not possibly see an "incorrect" frequency? In one trillion?

Pre-Bayesian: OK, you can in principle see an *incorrect* frequency, but it'd be ever less *likely*!

Bayesian: Tell me once again, what does "likely" mean?

This is a cartoon of course, but it captures the essence and the futility of every such debate. It is better to admit at the outset that probability is a degree of belief, and deal with the world on its own terms as it coughs up its objects and events. What do we gain for our theoretical conceptions by saying that along with each actual event there is a ghostly spirit (its "objective probability", its "propensity", its "objective chance") gently nudging it to happen just as it did? Objects and events are enough by themselves.

Similarly for quantum mechanics. Here too, if ghostly spirits are imagined behind the actual events produced in quantum measurements, one is left with conceptual troubles to no end. The defining feature of Quantum Bayesianism (Caves et al. 2002; Fuchs 2002; Fuchs and Schack 2004; Caves et al. 2007; Fuchs 2010; Fuchs and Schack 2013) is that it says along the lines of de Finetti, "If this is going to be done right, we must go this far." Specifically, there can be no such thing as a right and true quantum state, if such is thought of as defined by criteria *external* to the agent making the assignment: quantum states must instead be like personalist, Bayesian probabilities.

The direct connection between the two foundational issues is this. Quantum states, through the Born Rule, can be used to calculate probabilities. Conversely, if one assigns probabilities for the outcomes of a well-selected set of measurements, then this is mathematically equivalent to making the quantum-state assignment itself. The two kinds of assignments determine each other uniquely. Just think of a spin-$\frac{1}{2}$ system. If one has elicited one's degrees of belief for the outcomes of a σ_x measurement, and similarly one's degrees of belief for the outcomes of σ_y and σ_z measurements, then this is the same as specifying a quantum state itself: for if one knows the quantum state's projections onto three independent axes, then that uniquely determines a Bloch vector, and hence a quantum state. Something similar is true of all quantum systems of all sizes and dimensionality. There is no mathematical fact embedded in a quantum state ρ that is not embedded in an appropriately chosen set of probabilities. Thus generally, if probabilities are personal in the Bayesian sense, then so too must be quantum states.

What this buys interpretatively, beside airtight consistency with the best understanding of probability theory, is that it gives each quantum state a home. Indeed, a home localized in space and time – namely, the physical site of the agent who assigns it! By this method, one expels once and for all the fear that quantum mechanics leads to "spooky action at a distance", and expels as well any hint of a problem with "Wigner's friend". It does this because it removes the very last trace of confusion over whether quantum states might still be objective, agent-independent, physical properties.

The innovation here is that, for most of the history of efforts to take an informational point of view about quantum states, the supporters of the idea have tried to have it both ways: that on the one hand quantum states are not real physical properties, yet on the other there is a right quantum state independent of the agent after all. For instance, one hears things like, "The *right* quantum state is the one the agent should adopt if he had all the information." The tension in these two desires leaves their holders open to attack on both flanks and general confusion all around.

Take first instantaneous action at a distance – the distaste of this idea is often one of the strongest motivations for those seeking to take an informational stance on quantum states. Without the protection of truly personal quantum-state assignments, action at a distance is there as doggedly as it ever was. And things only get worse with "Wigner's friend" if one insists there be a *right* quantum state. As it turns out, the method of mending this conundrum displays one of the most crucial ingredients of QBism. Let us put it in plain sight.

"Wigner's friend" is the story of two agents, Wigner and his friend, and one quantum system – the only deviation we make from a more common presentation is that we put the story in informational terms. It starts off with the friend and Wigner having a conversation: suppose they both agree that some quantum state $|\psi\rangle$ captures their mutual beliefs about the quantum system. Furthermore suppose they agree that at a specified time the friend will make a measurement on the system of some observable (outcomes $i = 1, \ldots, d$). Finally, they both note that if the friend gets outcome i, he will (and should) update his beliefs about the system to some new quantum state $|i\rangle$. There the conversation ends and the action begins: Wigner walks away and turns his back to his friend and the supposed measurement. Time passes to some point beyond when the measurement should have taken place.

What now is the "correct" quantum state each agent should have assigned to the quantum system? We have already concurred that the friend will and should assign some $|i\rangle$. But what of Wigner? If he were to consistently dip into his mesh of beliefs, he would very likely treat his friend as a quantum system like any other: one with some initial quantum state ρ capturing his (Wigner's) beliefs of *it* (the friend), along with a linear evolution operator U to adjust those beliefs with the flow of time.[1] Suppose this quantum state includes Wigner's beliefs about everything he assesses to be interacting with his friend – in old parlance, suppose Wigner treats his friend as an isolated system. From this perspective, before any further interaction between himself and the friend or the other system, the quantum state Wigner would assign for the two together would be $U\big(\rho \otimes |\psi\rangle\langle\psi|\big)U^\dagger$ – most generally an entangled quantum state. The state of the system itself for Wigner would be gotten from this larger state by a partial trace operation; in any case, it will not be an $|i\rangle$.

Does this make Wigner's new state assignment incorrect? After all, "if he had all the information" (i.e., all the facts of the world) wouldn't that include knowing the friend's measurement outcome? Since the friend should assign some $|i\rangle$, shouldn't Wigner himself (if he had all the information)? Or is it the friend who is incorrect? For if the friend had "all the information", wouldn't he say that he is neglecting that Wigner could put the system and himself into the quantum computational equivalent of an iron lung and forcefully reverse the so-called measurement? I.e., Wigner, if he were sufficiently sophisticated, should be able to force

$$U\big(\rho \otimes |\psi\rangle\langle\psi|\big)U^\dagger \quad \longrightarrow \quad \rho \otimes |\psi\rangle\langle\psi| \; . \tag{1}$$

[1] For an explanation of the status of unitary operations from the QBist perspective, as personal judgments directly analogous to quantum states themselves, see Fuchs (2002) and Fuchs and Schack (2004).

And so the back and forth goes. Who has the *right* state of information? The conundrums simply get too heavy if one tries to hold to an agent-independent notion of correctness for otherwise personalistic quantum states. QBism dispels these and similar difficulties of the "aha, caught you!" variety by being conscientiously forthright. *Whose information?* "Mine!" *Information about what?* "The consequences (for *me*) of *my* actions upon the physical system!" It's all "I-I-me-me mine", as the Beatles sang.

The answer to the first question surely comes as no surprise by now, but why on earth the answer for the second? "It's like watching a QBist shoot himself in the foot", a friend once said. Why something so egocentric, anthropocentric, psychology-laden, myopic, and positivistic (we've heard any number of expletives) as *the consequences (for me) of my actions upon the system*? Why not simply say something neutral like "the outcomes of measurements"? To the uninitiated, our answer for *Information about what?* surely appears to be a cowardly, unnecessary retreat from realism. But it is the opposite. The answer we give is the very injunction that keeps the potentially conflicting statements of Wigner and his friend in check, at the same time as giving each agent a hook to the external world in spite of QBism's egocentric quantum states.

For QBists, the real world, the one both agents are embedded in – with its objects and events – is taken for granted. What is not taken for granted is each agent's access to the parts of it he has not touched. Wigner holds two thoughts in his head: (1) that his friend interacted with a quantum system, eliciting some consequence of the interaction for himself, and (2) after the specified time, for any of Wigner's own further interactions with his friend or system or both, he ought to gamble upon their consequences according to $U\left(\rho \otimes |\psi\rangle\langle\psi|\right)U^{\dagger}$. One statement refers to the friend's potential experiences, and one refers to Wigner's own. So long as it is kept clear that $U\left(\rho \otimes |\psi\rangle\langle\psi|\right)U^{\dagger}$ refers to the latter – how Wigner should gamble upon the things that might happen to him – making no statement whatsoever about the former, there is no conflict. The world is filled with all the same things it was before quantum theory came along, like each of our experiences, that rock and that tree, and all the other things under the sun; it is just that quantum theory provides a calculus for gambling on each agent's own experiences – it doesn't give anything else than that. It certainly doesn't give one agent the ability to conceptually pierce the other agent's personal experience. It is true that with enough effort Wigner could enact Eq. (1), causing him to predict that his friend will have amnesia to any future questions on his old measurement results. But we always knew Wigner could do that – a mallet to the head would have been good enough.

The key point is that quantum theory, from this light, takes nothing away from the usual world of common experience we already know. It only *adds*. At the very least it gives each agent an extra tool with which to navigate the world. More than that, the tool is here for a reason. QBism says when an agent reaches out and touches a quantum system – when he performs a *quantum measurement* – that process gives rise to birth in a nearly literal sense. With the action of the agent upon the system, the no-go theorems of Bell and Kochen-Specker assert that something new comes into the world that wasn't there previously: it is the "outcome", the unpredictable

consequence for the very agent who took the action. John Wheeler said it this way (Wheeler 1982), and we follow suit, "Each elementary quantum phenomenon is an elementary act of 'fact creation'."

With this much, QBism has a story to tell on both quantum *states* and quantum *measurements*, but what of quantum *theory* as a whole? The answer is found in taking it as a *universal* single-user theory in much the same way that Bayesian probability theory is. It is a users' manual that *any* agent can pick up and use to help make wiser decisions in this world of inherent uncertainty.[2] To say it in a more poignant way: in my case, it is a world in which *I* am forced to be uncertain about the consequences of most of *my* actions; and in your case, it is a world in which *you* are forced to be uncertain about the consequences of most of *your* actions. "And what of God's case? What is it for him?" Trying to give *him* a quantum state was what caused this trouble in the first place! In a quantum mechanics with the understanding that each instance of its use is strictly single-user – "My measurement outcomes happen right here, to me, and I am talking about my uncertainty of them." – there is no room for most of the standard, year-after-year quantum mysteries.

The only substantive *conceptual* issue left before synthesizing a final vaccine is whether quantum mechanics is obligated to derive the notion of agent for whose aid the theory was built in the first place? The answer comes from turning the tables: thinking of probability theory in the personalist Bayesian way, as an extension of formal logic, would one ever imagine that the notion of an agent, the user of the theory, could be derived out of its conceptual apparatus? Clearly not. How could you possibly get flesh and bones out of a calculus for making wise decisions? The logician and the logic he uses are two different substances – they live in conceptual categories worlds apart. One is in the stuff of the physical world, and one is somewhere nearer to Plato's heaven of ideal forms. Look as one might in a probability textbook for the ingredients to reconstruct the reader himself, one will never find them. So too, QBism says of quantum theory.

With this we finally pin down the precise way in which quantum theory is "different in character from any physical theory posed before." For QBism, quantum theory is not something *outside* probability theory – it is not a picture of the world as it is, as say Einstein's program of a unified field theory hoped to be – but rather it is

[2]Most of the time one sees Bayesian probabilities characterized as measures of ignorance or imperfect knowledge. But that description carries with it a metaphysical commitment that is not at all necessary for the personalist Bayesian, where probability theory is an extension of logic. Imperfect knowledge? It sounds like something that, at least in imagination, could be perfected, making all probabilities zero or one – one uses probabilities only because one does not know the true, pre-existing state of affairs. Language like this, the reader will notice, is never used in this paper. All that matters for a personalist Bayesian is that there is *uncertainty* for whatever reason. There might be uncertainty because there is ignorance of a true state of affairs, but there might be uncertainty because the world itself does not yet know what it will give – i.e., there is an objective indeterminism. As will be argued in later sections, QBism finds its happiest spot in an unflinching combination of "subjective probability" with "objective indeterminism."

an *addition* to probability theory itself. As probability theory is a *normative* theory, not saying what one *must* believe, but offering rules of consistency an agent should strive to satisfy within his overall mesh of beliefs, so it is the case with quantum theory.

To take this substance into one's mindset is all the vaccination one needs against the threat that quantum theory carries something viral for theoretical physics as a whole. A healthy body is made healthier still. For with this protection, we are for the first time in a position to ask, with eyes wide open to what the answer could not be, *just what after all is the world made of?* Far from being the last word on quantum theory, QBism is the start of a great adventure.

4 Hilbert-Space Dimension as a Universal Capacity

A common accusation heard is that QBism leads straight away to solipsism, "the belief that all reality is just one's imagining of reality, and that one's self is the only thing that exists." The accusation goes that, if a quantum state $|\psi\rangle$ only represents the degrees of belief held by some agent then the agent's beliefs must be the source of the universe. The universe could not exist without him: this being such a ridiculous idea, QBism is dismissed out of hand, *reductio ad absurdum*. It is so hard for the QBist to understand how anyone could think this (it being the antithesis of everything in his worldview) that a little of our own Latin comes to mind: *non sequitur*.

A fairer-minded assessment is that the accusation springs from our opponents "hearing" much of what we do say, but interpreting it in terms drawn from a particular conception of what physical theories *always ought to be*: attempts to directly represent (map, picture, copy, correspond to, correlate with) the *universe* – with "universe" here thought of as a static, timeless block that just *is*. From such a "representationalist" point of view, *if* **(a)** quantum theory is a proper physical theory, **(b)** its essential theoretical objects are quantum states, and **(c)** quantum states are states of belief, *then* the universe that "just is" corresponds to a state of belief. Solipsism on a stick, one might say.

QBism sidesteps the poisoned dart by asserting that quantum theory is just not a physical theory in the sense the accusers want it to be. Rather it is an addition to personal, Bayesian, normative probability theory. Its normative rules for connecting probabilities (personal judgments) were developed in light of the *character of the world*, but there is no sense in which the quantum state itself represents (pictures, copies, corresponds to, correlates with) a part or a whole of the external world, much less a world that *just is*. In fact the very character of the theory points to the inadequacy of the representationalist program when attempted on the particular world we live in.

QBism does not argue that representationalism must be wrong always and in all possible worlds (perhaps because of some internal inconsistency). Representationalism may well be true in this or that setting – we take no stand on

the matter. We only know that for nearly 90 years quantum theory has been actively resistant to representationalist efforts on *its* behalf. This suggests that it might be worth exploring some philosophies upon which physics rarely sets foot. Physics of course should never be constrained by any one philosophy (history shows it nearly always lethal), but it does not hurt to get ideas and insights from every source one can. If one were to sweep the philosophical literature for schools of thought representative of what QBism actually is about, it is not solipsism one will find, but nonreductionism (Dupré 1993; Cartwright 1999), (radical) metaphysical pluralism (James 1996a; Wahl 1925), empiricism (James 1940, 1996b), indeterminism and meliorism[3] (James 1884), and above all pragmatism (Thayer 1981).

A form of nonreductionism can already be seen in play in our answer to whether the notion of agent should be derivable from the quantum formalism itself. We say that it cannot be and it should not be, and to believe otherwise is to misunderstand the subject matter of quantum theory. But nonreductionism also goes hand in hand with the idea that there is real particularity and "interiority" in the world. Think again of the "I-I-me-me mine" feature that shields QBism from inconsistency in the "Wigner's friend" scenario. When Wigner turns his back to his friend's interaction with the system, that piece of reality – Bohr might call it a "phenomenon" – is hermetically sealed from him. It has an inside, a vitality that he takes no part in until he again interacts with one or both relevant pieces of it. *With respect to Wigner*, it is a bit like a universe unto itself.

If one seeks the essence of indeterminism in quantum mechanics, there may be no example more directly illustrative of it than "Wigner's friend". For it expresses to a tee William James's notion of indeterminism (James 1884, p. 145):

> [Chance] is a purely negative and relative term, giving us no information about that of which it is predicated, except that it happens to be disconnected with something else – not controlled, secured, or necessitated by other things in advance of its own actual presence. . . . What I say is that it tells us nothing about what a thing may be in itself to call it 'chance.' . . . All you mean by calling it 'chance' is that this is not guaranteed, that it may also fall out otherwise. For the system of other things has no positive hold on the chance-thing. Its origin is in a certain fashion negative: it escapes, and says, Hands off! coming, when it comes, as a free gift, or not at all.
>
> This negativeness, however, and this opacity of the chance-thing when thus considered *ab extra*, or from the point of view of previous things or distant things, do not preclude its having any amount of positiveness and luminosity from within, and at its own place and moment. All that its chance-character asserts about it is that there is something in it really of its own, something that is not the unconditional property of the whole. If the whole wants this property, the whole must wait till it can get it, if it be a matter of chance. That the universe may actually be a sort of joint-stock society of this sort, in which the sharers have both limited liabilities and limited powers, is of course a simple and conceivable notion.

[3] Strictly speaking, meliorism is the doctrine "that humans can, through their interference with processes that would otherwise be natural, produce an outcome which is an improvement over the aforementioned natural one." But we would be reluctant to take a stand on what "improvement" really means. So said, all we mean in the present essay by meliorism is that the world before the agent is malleable to some extent – that his actions really can change it.

The train of logic back to QBism is this. If James and our analysis of "Wigner's friend" are right, the universe is not *one* in a very rigid sense, but rather more truly a pluriverse.[4] To get some sense of what this can mean, it is useful to start by thinking about what it is not. A good example can be found by taking a solution to the vacuum Maxwell equations in some extended region of spacetime. Focus on a compact subregion and try to conceptually delete the solution within it, reconstructing it with some new set of values. It can't be done. The fields outside the region (including the boundary) uniquely determine the fields inside it. The interior of the region has no identity but that dictated by the rest of the world – it has no "interiority" of its own. The pluriverse conception says we'll have none of that. And so, for any agent immersed in this world there will always be uncertainty for what will happen upon his encounters with it.

What all this hints is that for QBism the proper way to think of our world is as the empiricist or radical metaphysical pluralist does. Let us launch into making this clearer, for that process more than anything will explain how QBism hopes to interpret Hilbert-space dimension.

The metaphysics of empiricism can be put like this. Everything experienced, everything experienceable, has no less an ontological status than anything else. You tell me of your experience, and I will say it is real, even a distinguished part of reality. A child awakens in the middle of the night frightened that there is a monster under her bed, one soon to reach up and steal her arm – that *we-would-call-imaginary* experience has no less a hold on onticity than a Higgs-boson detection event would if it were to occur at the fully operational LHC. They are of equal status from this point of view – they are equal elements in the filling out and making of reality. This is because the world of the empiricist is not a sparse world like the world of Democritus (*nothing but* atom and void) or Einstein (*nothing but* unchanging spacetime manifold equipped with this or that field), but a world overflowingly full of variety – a world whose details are beyond anything grammatical (rule-bound) expression can articulate.

Yet this is no statement that physics should give up, or that physics has no real role in coming to grips with the world. It is only a statement that physics should better understand its function. What is being aimed for here finds its crispest, clearest contrast in a statement of Richard Feynman (1965):

> If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generation of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic hypothesis (or the atomic fact) that all things are made of atoms – little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. . . . Everything is made of atoms. That is the key hypothesis.

[4]The term "pluriverse" is again a Jamesian one. He used it interchangeably with the word "multiverse", which he also invented. The latter however has been coopted by the Everettians, so we will strictly use only the term pluriverse.

The problem the imagery that usually lies behind the phrase "everything is made of". William James called it the great original sin of the rationalistic mind (James 1997, p. 246):

> Let me give the name of 'vicious abstractionism' to a way of using concepts which may be thus described: We conceive a concrete situation by singling out some salient or important feature in it, and classing it under that; then, instead of adding to its previous characters all the positive consequences which the new way of conceiving it may bring, we proceed to use our concept privatively; reducing the originally rich phenomenon to the naked suggestions of that name abstractly taken, treating it as a case of 'nothing but' that, concept, and acting as if all the other characters from out of which the concept is abstracted were expunged. Abstraction, functioning in this way, becomes a means of arrest far more than a means of advance in thought. It mutilates things; it creates difficulties and finds impossibilities; and more than half the trouble that metaphysicians and logicians give themselves over the paradoxes and dialectic puzzles of the universe may, I am convinced, be traced to this relatively simple source. *The viciously privative employment of abstract characters and class names* is, I am persuaded, one of the great original sins of the rationalistic mind.

What is being realized through QBism's peculiar way of looking at things is that physics *actually can be done* without any accompanying vicious abstractionism. You do physics as you have always done it, but you throw away the idea "everything is made of [Essence X]" before even starting.

Physics – in the right mindset – is not about identifying the bricks with which nature is made, but about identifying what is *common to* the largest range of phenomena it can get its hands on. The idea is not difficult once one gets used to thinking in these terms. Carbon? The old answer would go that it is *nothing but* a building block that combines with other elements according to the following rules, blah, blah, blah. The new answer is that carbon is a *characteristic* common to diamonds, pencil leads, deoxyribonucleic acid, burnt pancakes, the space between stars, the emissions of Ford pick-up trucks, and so on – the list is as unending as the world is itself. For, carbon is also a characteristic common to this diamond and this diamond and this diamond and this. But a flawless diamond and a purified zirconium crystal, no matter how carefully crafted, have no such characteristic in common: carbon is not a *universal* characteristic of all phenomena. The aim of physics is to find characteristics that apply to as much of the world in its varied fullness as possible. However, those common characteristics are hardly what the world is made of – the world instead is made of this and this and this. The world is constructed of every particular there is and every way of carving up every particular there is.

An unparalleled example of how physics operates in such a world can be found by looking to Newton's law of universal gravitation. What did Newton really find? Would he be considered a great physicist in this day when every news magazine presents the most cherished goal of physics to be a Theory of Everything? For the law of universal gravitation is hardly that! Instead, it *merely* says that every body in the universe tries to accelerate every other body toward itself at a rate proportional to its own mass and inversely proportional to the squared distance between them. Beyond that, the law says nothing else particular of objects, and it would have been a rare thinker in Newton's time, if any at all, who would have imagined that all the

complexities of the world could be derived from that limited law. Yet there is no doubt that Newton was one of the greatest physicists of all time. He did not give a theory of everything, but a Theory of One Aspect of Everything. And only the tiniest fraction of physicists of any variety have ever worn a badge of that more modest kind. It is as H. C. von Baeyer wrote in one of his books (von Baeyer 2009),

> Great revolutionaries don't stop at half measures if they can go all the way. For Newton this meant an almost unimaginable widening of the scope of his new-found law. Not only Earth, Sun, and planets attract objects in their vicinity, he conjectured, but all objects, no matter how large or small, attract all other objects, no matter how far distant. It was a proposition of almost reckless boldness, and it changed the way we perceive the world.

Finding a theory of "merely" one aspect of everything is hardly something to be ashamed of: it is the loftiest achievement physics can have in a living, breathing nonreductionist world.

Which leads us back to Hilbert space. Quantum theory – that user's manual for decision-making agents immersed in a world of *some* yet to be fully identified character – makes a statement about the world to the extent that it identifies a quality common to all the world's pieces. QBism says the quantum state is not one of those qualities. But of Hilbert spaces themselves, particularly their distinguishing characteristic one from the other, *dimension*, QBism carries no such grudge. Dimension is something one posits for a body or a piece of the world, much like one posits a mass for it in the Newtonian theory. Dimension is something a body holds all by itself, regardless of what an agent thinks of it.

The claim here is that quantum mechanics, when it came into existence, implicitly recognized a previously unnoticed capacity inherent in all matter – call it *quantum dimension*. In one manifestation, it is the fuel upon which quantum computation runs (Fuchs 2004; Blume-Kohout et al. 2002). In another it is the raw irritability of a quantum system to being eavesdropped upon Fuchs and Schack (2004).

When quantum mechanics was discovered, something was *added* to matter in our conception of it. Think of the apple that inspired Newton to his law. With its discovery the color, taste, and texture of the apple didn't disappear; the law of universal gravitation didn't reduce the apple privatively to *just* gravitational mass. Instead, the apple was at least everything it was before, but afterward even more – for instance, it became known to have something in common with the moon. A modern-day Cavendish would be able to literally measure the further attraction an apple imparts to a child already hungry to pick it from the tree. So similarly with Hilbert-space dimension. Those diamonds we have used to illustrate the idea of nonreductionism, in very careful conditions, could be used as components in a quantum computer (Prawer and Greentree 2008). Diamonds have among their many properties something not envisioned before quantum mechanics – that they could be a source of relatively accessible Hilbert space dimension and as such have this much in common with any number of other proposed implementations of quantum computing. Diamonds not only have something in common with the moon, but now with the ion-trap quantum-computer prototypes around the world.

Diamondness is not something to be derived from quantum mechanics. It is that quantum mechanics is something we *add* to the repertoire of things we already say of diamonds, to the things we do with them and the ways we admire them. This is a very powerful realization: for diamonds already valuable, become ever more so as their qualities compound. And saying more of them, not less of them as is the goal of all reductionism, has the power to suggest all kinds of variations on the theme. For instance, thinking in quantum mechanical terms might suggest a technique for making "purer diamonds". Though to an empiricist this phrase means not at all what it means to a reductionist. It means that these similar things called diamonds can suggest exotic variations of the original objects with various pinpointed properties this way or that. Purer diamond is not *more* of what it already was in nature. It is a new species, with traits of its parents to be sure, but nonetheless stand-alone, like a new breed of dog.

To the reductionist, of course, this seems exactly backwards. But then, it is the reductionist who must live with a seemingly infinite supply of conundrums arising from quantum mechanics. It is the reductionist who must live in a state of arrest, rather than moving on to the next stage of physics. Take a problem that has been a large theme of the quantum foundations meetings for the last 30 years. To put it in a commonly heard question, "Why does the world look classical if it actually operates according to quantum mechanics?" The touted mystery is that we never "see" quantum superposition and entanglement in our everyday experience.

The real issue is this. The expectation of the quantum-to-classical transitionists is that quantum theory is at the bottom of things, and "the classical world of our experience" is something to be derived out of it. QBism says "No. Experience is neither classical nor quantum. Experience is experience with a richness that classical physics of any variety could not remotely grasp." Quantum mechanics is something put on top of raw, unreflected experience. It is additive to it, suggesting wholly new types of experience, while never invalidating the old. To the question, "Why has no one ever *seen* superposition or entanglement in diamond before?", the QBist replies: it is simply because before recent technologies and very controlled conditions, as well as lots of refined analysis and thinking, no one had ever mustered a mesh of beliefs relevant to such a range of interactions (factual and counterfactual) with diamonds. No one had ever been in a position to adopt the extra normative constraints required by the Born Rule. For QBism, it is not the emergence of classicality that needs to be explained, but the emergence of our new ways of manipulating, controlling, and interacting with matter that do.

In this sense, QBism declares the quantum-to-classical research program unnecessary (and actually obstructive) in a way not so dissimilar to the way Bohr's 1913 model of the hydrogen atom declared another research program unnecessary (and actually obstructive). Bohr's great achievement above all the other physicists of his day was in being the first to say, "Enough! I shall not give a mechanistic explanation for these spectra we see. Here is a way to think of them with no mechanism." The important question is how matter can be coaxed to do new things. It is in the ways the world yields to our desires, and the ways it refuses to, that we learn the depths of its character.

5 The Future

There is so much still to do with QBism. So far we have only given the faintest hint of how QBism should be mounted onto a larger empiricism. It will be noticed that QBism has been quite generous in treating agents as physical objects when needed. "I contemplate you as an agent when discussing your experience, but I contemplate you as a physical system before me when discussing my own." Our solution to "Wigner's friend" is the great example of this. Precisely because of this, however, QBism knows that its story cannot end as a story of gambling agents – that is only where it starts. Agency, for sure, is not a derivable concept as the reductionists and vicious abstractionists would have it, but QBism, like all of science, should strive for a Copernican principle whenever possible. We have learned so far from quantum theory that before an agent the world is really malleable and ready through their intercourse to give birth. Why would it not be so for every two parts of the world? And this newly defined valence, quantum dimension, might it not be a measure of a system's potential for creation when it comes into relationship with those other parts?

It is a large research program whose outline is just taking shape. It hints of a world, a pluriverse, that consists of an all-pervading "pure experience", as William James called it.[5] Or, as John Wheeler put it in the form of a question (Wheeler 1982),

> It is difficult to escape asking a challenging question. Is the entirety of existence, rather than being built on particles or fields of force or multidimensional geometry, built upon billions upon billions of elementary quantum phenomena, those elementary acts of 'observer-participancy', those most ethereal of all the entities that have been forced upon us by the progress of science?

Expanding this notion, making it technical, and trying to weave its insights into worldview is the better part of future work. Quantum states, QBism declares, are not the stuff of the world, but quantum *measurement* might be. Might a one-day future Shakespeare write with honesty,

> Our revels are now ended. These our actors,
> As I foretold you, were all spirits and
> Are melted into air, into thin air ...
>
> We are such stuff as
> quantum measurement is made on.

[5] Aside from James's originals James (1996a,b), further reading on this concept and related subjects can be found in Lamberth (1999), Taylor and Wozniak (1996), Wild (1969), and Banks (2003).

References

Banks, E.C. 2003. *Ernst Mach's world elements: A study in natural philosophy*. Dordrecht: Kluwer.

Bell, J.S. 1990. Against 'measurement'. *Physics World* 3: 33.

Bernardo, J.M., and A.F.M. Smith 1994. *Bayesian theory*. Chichester: Wiley.

Blume-Kohout, R., C.M. Caves, and I.H. Deutsch. 2002. Climbing mount scalable: Physical-resource requirements for a scalable quantum computer. *Foundations of Physics* 32: 1641.

Cartwright, N. 1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Caves, C.M., C.A. Fuchs, and R. Schack. 2002. Quantum probabilities as Bayesian probabilities. *Physical Review A* 65: 022305.

Caves, C.M., C.A. Fuchs, and R. Schack. 2007. Subjective probability and quantum certainty. *Studies in History and Philosophy of Modern Physics* 38: 255.

de Finetti, B. 1990. *Theory of probability*. New York: Wiley.

Dupré, J. 1993. *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge: Harvard University Press.

Feynman, R.P. 1965. *The character of physical law*. Cambridge: MIT Press.

Fuchs, C.A. 2002. Quantum mechanics as quantum information (and only a little more). In *Quantum theory: Reconsideration of foundations*, ed. A. Khrennikov, 463–543. Växjö: Växjö University Press. `arXiv:quant-ph/0205039`.

Fuchs, C.A. 2004. On the quantumness of a hilbert space. *Quantum Information and Computation* 4: 467.

Fuchs, C.A. 2010. *Coming of age with quantum information*. Cambridge: Cambridge University Press.

Fuchs, C.A., and R. Schack. 2004. Unknown quantum states and operations, a Bayesian view. In *Quantum estimation theory*, ed. M.G.A. Paris and J. Řeháček, 151–190. Berlin: Springer.

Fuchs, C.A., and R. Schack. 2013, to appear. Quantum-bayesian coherence. *Reviews of Modern Physics*. `arXiv:1301.3274v1`.

Hartle, J.B. 1968. Quantum mechanics of individual systems. *American Journal of Physics* 36: 704.

James, W. 1884. *The will to believe and other essays in popular philosophy*; *Human immortality* – Both books bound as one. New York: Dover.

James, W. 1940. *Some problems of philosophy*. London: Longmans, Green, and Co.

James, W. 1996a. *A pluralistic universe*. Lincoln: University of Nebraska Press.

James, W. 1996b. *Essays in radical empiricism*. Lincoln: University of Nebraska Press.

James, W. 1997. *The meaning of truth*. Amherst: Prometheus Books.

Lamberth, D.C. 1999. *William James and the metaphysics of experience*. Cambridge: Cambridge University Press.

Lindley, D.V. 2006. *Understanding uncertainty*. Hoboken: Wiley.

Prawer, S., and A.D. Greentree. 2008. Diamond for quantum computing. *Science* 320: 1601.

Spekkens, R.W. 2007. Evidence for the epistemic view of quantum states: A toy theory. *Physical Review A* 75: 032110.

Taylor, E., and R.H. Wozniak, eds. 1996. *Pure experience: The response to William James*. Bristol: Thoemmes Press.

von Baeyer, H.C. 2009. *Petite Leçons de Physique dans les Jardins de Paris*. Paris: Dunod.

Thayer, H.S. 1981. *Meaning and action: A critical history of pragmatism*. 2nd ed. Indianapolis: Hackett Publishing.

Wahl, J. 1925. *The pluralist philosophies of England and America*. Trans. F. Rothwell. London: Open Court.

Wheeler, J.A. 1982. Bohr, Einstein, and the strange lesson of the quantum. In *Mind in nature*, ed. R.Q. Elvee, 1–23. San Francisco: Harper & Row.

Wild, J. 1969. *The radical empiricism of William James*. Garden City: Doubleday.

A Critic Looks at QBism

Guido Bacciagaluppi

1 Introduction

"M. Braque est un jeune homme fort audacieux. [...] Il méprise la forme, réduit tout, sites et figures et maisons, à des schémas géométriques, à des cubes. Ne le raillons point, puisqu'il est de bonne foi. Et attendons."[1] Thus commented the French art critic Louis Vauxcelles on Braque's first one-man show in November 1908, thereby giving cubism its name. Substituting spheres and tetrahedra for cubes might be more appropriate if one wishes to apply the characterisation to qBism – the view of quantum mechanics and the quantum state developed by Chris Fuchs and co-workers (for a general reference see either the paper in this volume, or Fuchs 2010). In this note, I shall not comment on other possible analogies, nor shall I present an exhaustive critical review of qBism (for an excellent one, see Timpson 2008). I simply wish to air a couple of worries I think qBists ought to think about more, in a friendlier spirit than the one Braque and Picasso's paintings might have encountered 100 years ago.

As was mentioned in discussion, qBism ought perhaps to stand not for "quantum Bayesianism", for there are many kinds of Bayesians who would not recognise themselves in it, but for "quantum Brunism", after Bruno de Finetti, who championed the radical subjectivist or pragmatist view of probabilities. In Sect. 2,

[1] Mr Braque is a very bold young man. [...] He despises form, reduces everything, places and figures and houses, to geometric schemes, to cubes. Let us not rail him, since he is in good faith. And let us wait. (My translation from Vauxcelles 1908)

G. Bacciagaluppi (✉)
Department of Philosophy, University of Aberdeen, Old Brewery, High Street, AB24 3UB Aberdeen, Scotland, UK

Institut d'Histoire et de Philosophie des Sciences et des Techniques, CNRS, Paris 1, ENS, France
e-mail: g.bacciagaluppi@abdn.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_27,

thus, I shall briefly remind the reader just how radical this view is. In Sect. 3 I shall then express a number of worries about the quantum case. Finally, Sect. 4 will give a quick sketch of some alternative pragmatist options for interpreting quantum probabilities.[2]

2 Brunism, Classical and Quantum

First of all, what are subjective probabilities (aka credences), or what are they *for* in a pragmatist understanding of probability? They are strategies we adopt in order to navigate the world. Some may turn out to be more useful than others in practice, but that does not change the fact that they are subjective constructs.

Take for instance (classical) coin tossing. Subjective probabilities describe our expectations for outcomes of successive tosses (and guide our betting behaviour). We start off with certain priors; these are updated as we go along, reflecting past performance. But where do our original priors come from? Let us dispel the idea that they stem from some intuitive feel or mystic vision. They might, but they generally will not, and indeed need not. Especially in more complex situations, we can and will adopt a theoretical model to choose our priors. Such a model may involve a thorough analysis of the whole set-up, or be even quite crude and just make use of simple properties of the coin alone (e.g. weight distribution). In any case, it will involve consideration of *objective* (but non-probabilistic!) properties of the system under consideration.

If the observed frequencies match the ones determined through the model, it may be that the data "confirm" our theoretical model. This, however, does not mean that our strategy is any less subjective. Indeed, there is no *necessary connection* between, say, weight distribution and relative frequencies: for any possible sequence of results there is an initial condition that will produce it, irrespective of the details of the weight distribution of the coin. Nor is there for the "Brunist" any *compelling rational justification* in the sense of David Lewis (1980) for letting our credences depend on the weight distribution, only pragmatic criteria such as simplicity, past performance etc. "Confirmation" is only a reflection of such past performance, and bears no guarantee of future reliability.

One might object that de Finetti himself showed that if one's priors for sequences of results are exchangeable, the posteriors will converge (with subjective

[2]With the addition of Sect. 3.1, the material in this note was presented at the Bertinoro conference. My thanks go to Maria Carla Galavotti, Raffaella Campaner, Beatrice Collina and all the colleagues in the ESF network who organised the conference. But most of all I wish to thank Chris Fuchs for the pleasure of endless discussions of quantum Bayesianism over the years, and for removing the prejudice in my mind that a subjective interpretation of quantum probabilities must be a non-starter.

probability 1). This is surely a sign of tracking something "out there" (or so the objection goes).[3] As long as we are dealing with finite samples, however, any apparent convergence of one's probability assignments (one's own or intersubjectively) again merely reflects past performance. And the expectation of future convergence depends crucially on the subjective assumption of exchangeability. Exchangeability is an assumption about the structure of one's subjective priors, and is entirely independent of any objective behaviour of the system under consideration.

Even this more technical objection, thus, does not alter the fact that according to de Finetti there is no sense in which our probability judgements are *right or wrong*. As de Finetti very graphically expresses,

PROBABILITIES DO NOT EXIST.

The same is true in qBism. Quantum probabilities according to qBism simply *are* subjective probabilities in the sense of de Finetti. They are strategies that we adopt in navigating an (admittedly) unexpectedly strange world. They may be impressively successful in terms of calculational power, past performance etc., but that makes them no less subjective than the theoretical models we might adopt for coin tossing (or indeed in classical statistical mechanics!).

What may strike one as strange, or at least unfamiliar, is applying de Finetti's ideas to a case which is generally thought to be genuinely *indeterministic*. There may be a price to pay if one does so, because if probabilities are not objective, then any use of them in describing the observed regularities in the world cannot be thought of as expressing law-like behaviour in any *necessary* sense. But if one is happy with a broadly Humean view of laws, this will not strike one as a disadvantage of the application of de Finetti's views to an indeterministic context.

Another point worth making explicitly (because it is not normally included in presentations of qBism), is that even in the quantum case one may adopt theoretical models for fixing one's priors. For instance, one might use techniques for selecting a Hamiltonian and calculating its ground state, and set one's subjective probabilities according to that. The data might then seem to "confirm" some particular choice of Hamiltonian. But according to the qBist there will be no necessity or rationally compelling reasons for using it to fix our credences, only pragmatic ones such as calculational power, past performance etc.

Fuchs generally talks of quantum states themselves as subjective (and this leads him also to a view of Hamiltonians as subjective). What I am saying instead is that one can take Hamiltonians to be objective properties of quantum systems, just like the weight distribution in a coin, and further as *non-probabilistic* properties, again just like weight distribution. Indeed, one can even think of quantum states (!) as objective properties of a system, stripping them of their customary probabilistic elements, and thinking for instance of the ground state "merely" as a state of definite energy. (After all, Heisenberg, Pauli, Schrödinger and others all had notions of

[3] Such a view seems to be suggested for instance by Greaves and Myrvold (2010).

quantum states *before* Born introduced his "statistical" interpretation of the wave function.[4]) The probabilistic association would attach to such a state exclusively in the way we use it to fix our credences. We shall return to the idea of objective non-probabilistic quantum states in Sect. 4.

3 Worries

3.1 EPR and Wigner's Friend

One worry I wish to air about qBism relates to one of its reputed major selling points, namely its "local" account of "nonlocal" EPR correlations. The qBist story goes like this. Consider Alice's credences about *Bob's* electron. Her initial credences might be equal to 1/2 for the result of any spin experiment she might perform on Bob's side. But if she performs first a spin measurement on her own side – say in direction x with result "up" –, she will then update her credences about further measurements on Bob's electron. In particular she will then believe with all her heart (credence 1) that if she performs a spin-x experiment on Bob's electron, this will give the result "down". What has changed, however, are simply *her own credences*. Similarly, Bob has credences about Alice's electron, which also change in an analogous manner, but although they are *about* Alice's electron, they are *his* credences, and (insofar as beliefs can be said to be located anywhere[5]) they are in his head. These two autonomous points of view can be then married together to form a composite picture of the EPR pair. (Thus, indeed, bringing out an analogy between qBism and cubism!)

The worry is about this marriage. QBism, just like "classical Brunism", presupposes that different agents be able to share data, an assumption that underlies the intersubjective agreement between different agents derivable from de Finetti's theorem. Different agents may become aware at different times of different pieces of data, but they can inspect each others' data and pool them together.

Indeed, suppose Alice and Bob both perform spin measurements in direction x, and then later meet to compare results. From Alice's point of view, her asking Bob his result is her own measurement of Bob's electron (which in the meantime has interacted with this further system called Bob), and her own measurement results are timelike related. But if she wishes to take Bob's own report seriously

[4]When Schrödinger introduced his wave functions, he clearly understood them as representing physical states of matter. But also Heisenberg and other advocates of matrix mechanics had a notion of stationary state, both preceding and distinct from Schrödinger's wave functions. Cf. Bacciagaluppi and Valentini (2009, ch. 3) and Bacciagaluppi (2008).

[5]Disregarding notions of extended cognition – which are presumably beside the point here.

as providing her with data that at the time of his measurement were available to him but not yet to her, then the mystery of perfectly correlated spacelike separated events returns.[6]

Wigner's friend (another case for which qBism claims to have a ready explanation) can be seen as a variant of the above. Bob performs an experiment in a lab that is isolated from Alice (perhaps because Alice and Bob are spacelike separated at the time). Alice can later perform a measurement on the content of Bob's lab (either repeating Bob's measurement or asking him for a report). If she does repeat Bob's measurement, the result she observes coincides with the result of the earlier measurement as reported by Bob, again suggesting that she should take the report seriously as describing objective data that were not yet available to her. Unlike the EPR case, this is not particularly puzzling. But in the Wigner's friend scenario, we are invited to consider also the case in which Alice performs instead an interference experiment on the entire contents of Bob's lab, and thereby "quantum erases" Bob's result. In qBist terms, this could be understood merely as Alice performing some manipulation that leads her to change her own credences about the results of her asking Bob what he has seen. She now expects from Bob not some or other report of a definite result, but a definite report of not having performed the experiment. But this description misses out on the fact that Alice's manipulation has in fact obliterated also Bob's piece of data and any memory that Bob had of it (unless, that is, we assume that Bob did not really possess any such piece of data in the first place).

Thus, if we believe that data obtained by different agents are equally objective, thus understanding "pooling of data" literally, we have problems. There are no such problems once the data have been pooled together, but we have two puzzling cases in situations where Bob's data are not yet available to Alice. In the EPR case, qBism remains silent on why Alice and Bob's data should be correlated, and in the Wigner's friend case, it remains silent on how Alice can erase Bob's data. The choice for qBists seems to be between: (a) providing us with a further story about data and/or agents *themselves*, rather than just strategies for how agents update their credences in the face of new data; and (b) some kind of solipsism or radical relativism, in which we care only about single individuals' credences, and not about whether and how they ought to mesh.

[6]The situation is quite analogous to that of collapse on the forward light cone. If technically feasible, a theory in which the collapse of the quantum state takes place along the forward light cone of a triggering event would be manifestly Lorentz-invariant. However, in the case of an EPR pair, it would leave unexplained how space-like separated collapses are correlated so as to match up on the overlap of their future light cones. Cf. e.g. the discussion in Bacciagaluppi (2010).

3.2 Hidden Constraints on Probability Assignments

My main worry, however, runs deeper in the conceptual foundations of qBism. A central idea of qBism is that the view is not a modification of but an *addition to* Bayesian coherence. This addition is equally normative, but rather than being rational in origin, it is empirically motivated. It is essentially contained in the formula

$$Q(D_j) = (d+1)\sum_{i=1}^{d^2} P(H_i)P(D_j|H_i) - 1, \tag{1}$$

which constrains the relation between probabilities in certain pairs of actual and counterfactual situations. The actual situation is a measurement of the family of projections D_j, and the formula compares the probabilities $Q(D_j)$ with the probabilities $\sum_{i=1}^{d^2} P(H_i)P(D_j|H_i)$ that *would* have been obtained if a previous measurement of the 'fiducial SIC' with effects H_i had been performed (a generalisation of this formula holds if the D_j are effects).[7] This is, indeed, a situation on which Bayesian coherence is silent. The law of total probability

$$Q(D_j) = \sum_{i=1}^{d^2} P(H_i)P(D_j|H_i) \tag{2}$$

is clearly a prescription for relating the probabilities of two *actual* measurements.

However, I claim that formula (1) already presupposes very strong constraints on our subjective probability assignments. Indeed, the very idea of well-defined probabilities $Q(D_j)$ for projections (or more generally effects) D_j already embodies such strong constraints, even before we start comparing our probability assignments for actual measurements to our probability assignments for counterfactual measurements.

In order to substantiate this claim, let me recall some standard material. It is nowadays customary in quantum mechanics and quantum information to describe (general) measurements using the concepts of *operations* and of *POVM*s (positive-operator-valued measures).

Operations are families of transformations on the quantum states (thought of as transformations induced by "measurements"). Such transformations could for instance be of the form ("pure operation")

$$\rho \mapsto A_i \rho A_i^* \tag{3}$$

(with the right-hand side suitably renormalised), and each such transformation takes place with probability

[7] See below for the definition of an effect.

$$\mathrm{Tr}(A_i \rho A_i^*) = \mathrm{Tr}(\rho A_i^* A_i) \; . \tag{4}$$

For this expression to indeed define a probability distribution over the various possible transformations we must have:

$$\sum_i D_i := \sum_i A_i^* A_i = \mathbf{1} \tag{5}$$

(with $\mathbf{1}$ the identity operator). The thus defined operators D_i are so-called *effects*, i.e. they are positive (self-adjoint with positive spectrum) and with spectrum contained in the interval $[0, 1]$. The mapping from the indices i (or sets thereof) to the associated D_i (or sums thereof) is thus an effect-valued measure, also called positive-operator-valued measure (POVM).

One should note crucially that the probabilities associated with a transformation are fixed just by the corresponding POVM.

Further, one easily sees that a pure operation such as the above can always be implemented by coupling the system to an ancillary system,

$$|\psi\rangle \otimes |\varphi_0\rangle \mapsto \sum_i A_i |\psi\rangle \otimes |\varphi_i\rangle \tag{6}$$

for some orthonormal family $|\varphi_i\rangle$, and then collapsing onto the latter. Note that such coupling is indeed unitary because of (5). (This result, suitably generalised to all operations, is known as the Naimark dilation theorem.)

Here is a very familiar example.

Example 1 (von Neumann measurement). Let $A_i := |\psi_i\rangle\langle\psi_i|$ for some orthonormal basis. We implement it via

$$\sum_i \alpha_i |\psi_i\rangle \otimes |\varphi_0\rangle \mapsto \sum_i \alpha_i |\psi_i\rangle \otimes |\varphi_i\rangle \; , \tag{7}$$

and we have

$$A_i^* A_i \; = \; P_i := |\psi_i\rangle\langle\psi_i| \; , \tag{8}$$

so the corresponding POVM is projection-valued.

A von Neumann measurement, however, is not the only experimental procedure for measuring a projection-valued measure, as the following example shows.

Example 2 ("measurement of the second kind"). Let the $|\psi_i\rangle$ be as above, and let $B_i = |\psi_i'\rangle\langle\psi_i|$ for some arbitrary unit vectors $|\psi_i'\rangle$. We have

$$\sum_i \alpha_i |\psi_i\rangle \otimes |\varphi_0\rangle \mapsto \sum_i \alpha_i |\psi_i'\rangle \otimes |\varphi_i\rangle \; . \tag{9}$$

Note that

$$B_i^* B_i = |\psi_i\rangle\langle\psi_i'|\psi_i'\rangle\langle\psi_i| = P_i \ , \tag{10}$$

and the transformation is indeed associated to the *same* projection-valued measure as in Example 1. The difference is that the "collapsed" state of the system after the measurement is no longer an eigenstate of the measured observable (in the traditional sense of a self-adjoint operator with spectral measure defined by the P_i). The measurement is not "minimally disturbing".

How do these examples lead to a worry about qBism? Note that (1) contains probability assignments $Q(D_j)$ referring to measurements of POVMs *irrespective* of which transformations are used to implement them. Thus, it presupposes that we assign the same probabilities to the results of the two transformations in Examples 1 and 2, even though they correspond to *different lab procedures*. QBism is currently silent on why we have this constraint on our probability assignments. Of course, we can say it is empirically well-established, and we can derive it theoretically from the quantum mechanical theory of measurement *if* we apply the usual Born rule to the ancillary system in the model. But if (1) is meant to be a simple axiom embodying one of the main modifications of Bayesian coherence theory that are supposed to *lead* to quantum mechanics, it appears that it already presupposes a lot of the structure it is trying to explain.

The point can be made even more strikingly using a further example.

Example 3 (sequential von Neumann measurements). Concatenating two operations yields a further operation, e.g.

$$\rho \mapsto P_i \rho P_i \mapsto Q_j P_i \rho P_i Q_j \ . \tag{11}$$

Indeed, defining

$$A_{ij} := Q_j P_i \ , \tag{12}$$

we obtain a corresponding POVM:

$$\sum_{ij} A_{ij}^* A_{ij} = \sum_{ij} P_i Q_j P_i = \sum_i P_i = \mathbf{1} \ . \tag{13}$$

This POVM can be measured via two sequential von Neumann measurements. For instance, if the P_i and Q_j project onto spin-1/2 eigenstates in directions x and y, we can implement the composite POVM by letting an electron pass in sequence two Stern–Gerlach magnets at right angles to each other (two sequential interactions between the spin and spatial degrees of freedom of the electron) and then measuring on which *quadrant* of the screen the electron impinges.

But we can also implement it using a *single* interaction with an ancillary system, e.g. by defining

$$B_{ij} := \sqrt{P_i Q_j P_i} \ . \tag{14}$$

Indeed,

$$B_{ij}^* B_{ij} = (\sqrt{P_i Q_j P_i})^2 = P_i Q_j P_i = A_{ij}^* A_{ij} \ , \tag{15}$$

and the corresponding POVM is the same as in (13).

This is again a case of two totally different laboratory procedures that allow one to measure the same POVM, this time an effect-valued rather than projection-valued one.

We can again note that the qBist formula relating the probabilities we assign to results of (actual) measurements of effect-valued measures to the probabilities in the corresponding counterfactual situations (a fairly straightforward generalisation of (1)) presupposes that we assign the same probabilities also to these two procedures.

This presupposition, however, is even more suspect than in the case of Examples 1 and 2, because the procedure defined by (12) already implicitly contains the *collapse* of the state (or at least the effective collapse through decoherence of the spin state by the spatial degree of freedom along the first measured direction x), which even more strongly suggests that the strategy currently pursued within qBism for axiomatising and/or understanding quantum mechanics implicitly presupposes a lot of the structure it is trying to explain.

4 Other Pragmatist Alternatives

Whether or not qBism will turn out to provide a fully successful new framework for understanding quantum mechanics, it has already shattered a taboo: that of using a subjectivist approach to probabilities in quantum mechanics. The traditional view of course is that quantum probabilities are *the* paradigm of objective probabilities. The idea that a radical subjectivist approach *à la* de Finetti might be applied to the quantum case (which at least *prima facie* is truly indeterministic) used to be inconceivable.

The inconceivable having now been conceived, I wish to suggest in this section that in fact a subjectivist/pragmatist approach to probability can be applied in the context of just about *any* approach to quantum mechanics. In particular, it can be applied also to approaches that adopt an ontic view of the quantum state. As pointed out in Sect. 2, an ontic view of the quantum state might be adopted also within qBism, as long as the quantum state itself is not seen as a probabilistic entity (although this is not part of the usual presentations of qBism[8]). Other approaches to

[8] I cannot resist teasing Chris here by pointing out that *pace* the remarks in Fuchs (2010, pp. 24–25), the results of decoherence just scream out for an ontic interpretation of the quantum state (so much classical-like structure within the quantum state that could be used to explain the classical

the foundations of quantum mechanics such as the Bohm theory, GRW theories, or the Everett theory explicitly take an ontic view of the quantum state (thus at least in principle also providing an answer to the question of the ontology of data and agents, cf. above Sect. 3.1).

I shall now sketch very briefly in turn how each of these can adopt a radical subjectivist view of probabilities, thus taking the quantum state as ontic but as non-probabilistic (whether or not it is customary for them to do so!).

4.1 Bohm

It is easiest to think of subjectivism about probabilities in the case of the Bohm theory. Indeed, the Bohm theory is a deterministic theory, and we are familiar with applying the classic de Finetti analysis to deterministic cases.

The Bohm theory (or de Broglie–Bohm theory, or pilot-wave theory) describes "classical" configurations evolving deterministically in a way fixed by the quantum wave function of the total system. The usual statistical predictions of quantum mechanics are recovered if one assumes that the configurations in an ensemble are distributed according to the usual quantum mechanical distribution (a condition which is preserved over time). The situation is very similar to that of classical statistical mechanics (with the difference that now the "equilibrium" distributions are time-dependent), and this analogy has been developed in considerable detail.[9]

Note that the quantum states in the Bohm theory are both ontic *and* non-probabilistic. They are the "pilot waves" of the theory. They acquire a probabilistic significance only if we adopt a strategy of choosing our subjective probabilities to be equal to the modulus squared of the wave function. While such a strategy is highly successful,[10] the quantum wave functions provide *no fundamental constraint* on initial positions, nor indeed on particle distributions in ensembles. This is evident from the fact that *non-equilibrium* pilot-wave theory is equally intelligible and may even have very interesting applications (Valentini 2010).

Just as in the case of the weight distribution of a coin – which turns out to be a reliable indicator of the statistical behaviour under repeated tossing only under certain "typical" conditions –, so the wave function describing an ensemble of particles (more precisely, the so-called effective wave function, i.e. the component of the wave function responsible for the motion) is a reliable indicator of the

world – *if only* we could avail ourselves to an ontic interpretation of the state). Cf. Bacciagaluppi (2012, esp. Sect. 3.5). The quantum-to-classical transition is here to stay, and qBism ought to incorporate it.

[9]For general references to the Bohm theory, see e.g. Bohm and Hiley (1993), Goldstein (2013), and Holland (1993).

[10]And can be justified using arguments analogous to those employed in classical statistical mechanics (see e.g. Valentini 1991; Dürr et al. 1992; Towler et al. 2011).

statistical behaviour of the particles only on the assumption of typicality. Thus in both cases a physical but non-probabilistic property associated with the system under consideration (and one that can be modelled theoretically) is used as a pragmatic short-cut for fixing our subjective probabilities for the behaviour of the system.

4.2 GRW

It is less familiar to think in terms of subjective probabilities in the case of spontaneous collapse theories, i.e. theories in which the Schrödinger evolution is modified in a way that reproduces the phenomenology of collapse. Such theories were shown to be viable by Ghirardi et al. (1986) and Pearle (1976, 1989), and are thus also know as GRW (or GRWP) theories.[11]

Since in GRW theories we have genuine indeterministic evolution of the quantum state, they are generally thought of in terms of objective collapse probabilities, much like "traditional" quantum mechanics. However, the idea that in the case of genuine indeterminism probabilities should be thought of as objective presupposes that a viable account of objective probabilities be given, e.g. in terms of frequencies or propensities, and both of these accounts suffer from more or less severe problems. The third strategy open to objectivists is to apply Lewis's (1980) "principal principle", i.e. to argue that there are compelling rational reasons for adopting a particular recipe to fix one's subjective probabilities. Perhaps some version of "Humean objective chances" can deliver on this, but the step to subjectivism might be very short in that case.[12]

The quantum state in GRW theories is clearly ontic, indeed at least *prima facie* provides the "stuff" the world is made of.[13] But as regards the probabilistic *evolution* of the state, we can adopt de Finetti's position, holding that there are no right or wrong probabilities about how the state evolves. We can take the GRW theory (a theoretical model of these probabilities) as a pragmatic recipe for fixing our *subjective* probabilities for the dynamical behaviour of the states (much as we take weight distributions to guide our expectations about the behaviour of tossed coins). And we can push the line that de Finetti's position is not only tenable in an indeterministic context, but that it may be even less artificial than others in the context of GRW theories.

[11]For an accessible reference, see Ghirardi (2011).

[12]On Humean objective chances, see e.g. Hoefer (2007), and for their application to GRW and for more general discussion of objective probabilities in GRW, see Frigg and Hoefer (2007).

[13]There is a debate about the most natural interpretation of spontaneous collapse theories: whether – in increasing order of resilience against the so-called "tails" problem – it is in terms of wave functions, matter density, or collapse events (so-called "flashes" or "hits"). For details see e.g. Ghirardi (2000), Allori et al. (2008), and Bacciagaluppi (2010).

4.3 Everett

The final and most interesting case is that of Everett.[14] As is well known, Everett takes the wave function of the universe seriously as providing the ontology of the theory, and the Schrödinger equation as providing its dynamics. Collapse is explained through the correlational structure of the universal wave function, whereby the quantum state *appears* to collapse to an internal observer (or whatever other system is "recording" collapse events). Each outcome of a collapse is equally real, *relative* to the corresponding component of the observer.

Modern-day Everettians refine Everett's original analysis of the correlational structure of the universal wave function through the use of decoherence theory (cf. e.g. Wallace 2010a; Bacciagaluppi 2012). However, there is a generalised perception of a problem in making sense of probabilities in the Everett theory, precisely because, say, in the context of a sequence of measurements on an ensemble of systems, *all* sequences of outcomes are actualised (in different "worlds" or "branches" of the universal wave function). Thus, not only are there no right or wrong probabilities, but probabilities would appear to make no sense at all (at least in the sense that we do not seem to be ignorant of what will be the outcome of a measurement[15]).

A breakthrough on this question was achieved not many years ago by David Deutsch (1999) and David Wallace (2007; 2010b), who adopted the Lewisian strategy sketched above and argued, first, that rational decision theory can be applied to the case of an Everettian agent located before a branching event (see also Greaves 2004; Greaves and Myrvold 2010), and, crucially, that *rationality constraints* on such an agent will force them to adopt the quantum probabilities for the results of the branching. Thus, quantum probabilities (at least insofar as they apply to such a decision-theoretic situation) are objective chances in the sense of Lewis. The approach based on the Deutsch–Wallace theorem appears to command quite a consensus among modern Everettians, but has also been the object of strong criticism.[16]

I wish to suggest that Everettians can avail themselves of an alternative *subjectivist* strategy, taking "branch weights" as guides for navigating a branching universe. The choice of branch weight as quantifying probability (or "typicality")

[14]Cf. also my comments in Bacciagaluppi (2013), which reviews the state-of-the-art volume on the Everett theory edited by Saunders et al. (2010). Everett's complete writings on quantum mechanics, together with a wealth of other original material, have been published and annotated by Barrett and Byrne (2012).

[15]Note that, as clearly shown by Vaidman (1996), we do have ignorance of the result of a measurement at least *after* the measurement has occurred and we are not yet aware of the result. After the branching induced by a measurement there is a genuine question about self-location.

[16]For a lively and representative sample of the literature, see the relevant contributions by Albert, Greaves and Myrvold, Kent, Price, Saunders, and Wallace, as well as the transcripts of the discussions, in Saunders et al. (2010). In particular, Price (2010) argues that agents may have *global* reasons on which to base their decisions, i.e. reasons other than the utilities of their successors. Such arguments of course undermine the idea that there should be compelling rational arguments for adopting the usual quantum probabilities in Everett.

can be pragmatically justified on the basis that it has performed well in the past, and on the basis of its being the "natural" measure on branches, e.g. because of Gleason's (1957) theorem or the Deutsch–Wallace theorem (this is Greaves and Myrvold's (2010) take on the latter), or because it is conserved by the dynamics analogously to the measure in classical statistical mechanics (this is in fact the justification proposed by Everett (see e.g. Barrett and Byrne 2012, pp. 274–275)). This reading of probability in Everett needs of course to be developed further, but would provide a particularly striking way of combining an ontic view of the quantum state with a subjectivist view of quantum probability.[17]

Such an application of "Brunism" to the quantum case would of course be much tamer than Chris Fuchs's usual emphasis on quantum states themselves being subjective. Stretching the metaphor, if we allow Everett's universal wave function to explain what agents and outcomes are in the first place, the heroic phase of "analytical qBism" will give way to a much tamer "synthetic qBism". That said – returning to the history of art – I have always preferred the analytic phase of *cubism* to the synthetic one!

References

Allori, V., S. Goldstein, R. Tumulka, and N. Zanghì. 2008. On the common structure of Bohmian mechanics and the Ghirardi–Rimini–Weber theory. *The British Journal for the Philosophy of Science* 59: 353–389.

Bacciagaluppi, G. 2008. The statistical interpretation according to Born and Heisenberg. In *HQ-1: Conference on the history of quantum physics*, MPIWG preprint series, vol. 350/II, ed. C. Joas, C. Lehner, and J. Renn, 269–288. Berlin: MPIWG. http://www.mpiwg-berlin.mpg.de/en/resources/preprints.html.

Bacciagaluppi, G. 2010. Collapse theories as beable theories. *Manuscrito* 33(1): 19–54, philSci-8876.

Bacciagaluppi, G. 2012. The role of decoherence in quantum mechanics. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta. Winter 2012 edn. http://plato.stanford.edu/archives/win2012/entries/qm-decoherence.

Bacciagaluppi, G. 2013. The many facets of Everett's many worlds. *Metascience* 22: 575–582.

Bacciagaluppi, G., and A. Valentini. 2009. *Quantum theory at the crossroads: Reconsidering the 1927 Solvay conference*. Cambridge: Cambridge University Press.

Barrett, J.A., and P. Byrne (eds.). 2012. *The Everett interpretation of quantum mechanics: Collected works 1955–1980 with commentary, by Hugh Everett III*. Princeton/Oxford: Princeton University Press.

Bohm, D., and B.J. Hiley. 1993. *The undivided universe: An ontological interpretation of quantum theory*. London: Routledge.

Deutsch, D. 1999. Quantum theory of probability and decisions. *Proceedings of the Royal Society of London A* 455: 3129–3137.

Dürr, D., S. Goldstein, and N. Zanghì. 1992. Quantum equilibrium and the origin of absolute uncertainty. *Journal of Statistical Physics* 67: 843–907.

[17] In further work, in particular with my graduate students, I hope to elaborate both on the analogy between Everett's view of probability in his own theory and in classical statistical mechanics – in particular on how it allows one to make statistical inferences in an Everettian universe –, and on the pragmatist reading of typicality, both in Everett and in classical statistical mechanics.

Frigg, R., and C. Hoefer. 2007. Probability in GRW theory. *Studies in History and Philosophy of Modern Physics* 38: 371–389.

Fuchs, C. 2010. QBism, the perimeter of quantum Bayesianism, arXiv:1003.5209.

Ghirardi, G.C. 2000. Local measurements of nonlocal observables and the relativistic reduction process. *Foundations of Physics* 30: 1337–1385.

Ghirardi, G.C. 2011. Collapse theories. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta, Winter 2011 edn. http://plato.stanford.edu/archives/win2011/entries/qm-collapse.

Ghirardi, G.C., A. Rimini, and T. Weber. 1986. Unified dynamics for microscopic and macroscopic systems. *Physical Review D* 34: 470–491.

Gleason, A. 1957. Measures on the closed subspaces of a Hilbert space. *Journal of Mathematics and Mechanics* 6: 885–893.

Goldstein, S. 2013. Bohmian mechanics. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta, Spring 2013 edn. http://plato.stanford.edu/archives/spr2013/entries/qm-bohm.

Greaves, H. 2004. Understanding Deutsch's probability in a deterministic multiverse. *Studies in History and Philosophy of Modern Physics* 35: 423–456.

Greaves, H., and W. Myrvold. 2010. Everett and evidence. In *Many worlds? Everett, quantum theory, and reality*, ed. S. Saunders, J. Barrett, A. Kent, and D. Wallace, 264–304. Oxford: Oxford University Press.

Hoefer, C. 2007. The third way on objective probability: A sceptic's guide to objective chance. *Mind* 116: 549–596.

Holland, P.R. 1993. *The quantum theory of motion: An account of the de Broglie-Bohm causal interpretation of quantum mechanics*. Cambridge: Cambridge University Press.

Lewis, D. 1980. A subjectivist's guide to objective chance. In *Studies in inductive logic and probability*, ed. R. C. Jeffrey, vol. II, 263–293. Berkeley: University of California Press. Reprinted in Philosophical papers, D. Lewis, vol. II, 159–213. Oxford: Oxford University Press.

Pearle, P. 1976. Reduction of the state vector by a nonlinear Schrödinger equation. *Physical Review D* 13: 857–868.

Pearle, P. 1989. Combining stochastic dynamical state-vector reduction with spontaneous localization. *Physical Review A* 39: 2277–2289.

Price, H. 2010. Decisions, decisions, decisions: Can Savage salvage Everettian probability? In *Many worlds? Everett, quantum theory, and reality*, ed. S. Saunders, J. Barrett, A. Kent, and D. Wallace, 369–390. Oxford: Oxford University Press.

Saunders, S., J. Barrett, A. Kent, and D. Wallace, eds. 2010. *Many worlds? Everett, quantum theory, and reality*. Oxford: Oxford University Press.

Timpson, C.G. 2008. Quantum Bayesianism: A study. *Studies in History and Philosophy of Modern Physics* 39: 579–609.

Towler, M.D., N.J. Russell, and A. Valentini. 2011. Timescales for dynamical relaxation to the Born rule. arXiv:1103.1589.

Vaidman, L. 1996. On schizophrenic experiences of the neutron or why we should believe in the many-worlds interpretation of quantum theory. arXiv:quant-ph/9609006.

Valentini, A. 1991. Signal-locality, uncertainty, and the subquantum H-theorem, I and II. *Physics Letters A* 156: 5–11; 158: 1–8.

Valentini, A. 2010. Inflationary cosmology as a probe of primordial quantum mechanics. *Physical Review D* 82: 063513. arXiv:0805.0163.

Vauxcelles, L. 1908. Exposition Braque. *Gil Blas* 14 Nov 1908. Quoted and reproduced in Worms de Romilly, N., and J. Laude (eds.). 1982. *Braque: le cubisme, fin 1907–1914*. Paris: Galerie Maeght, 11, 13.

Wallace, D. 2007. Quantum probability from subjective likelihood: Improving on Deutsch's proof of the probability rule. *Studies in History and Philosophy of Modern Physics* 38: 311–332.

Wallace, D. 2010a. Decoherence and ontology. In *Many worlds? Everett, quantum theory, and reality*, ed. S. Saunders, J. Barrett, A. Kent, and D. Wallace, 53–72. Oxford: Oxford University Press.

Wallace, D. 2010b. How to prove the Born rule. In *Many worlds? Everett, quantum theory, and reality*, ed. S. Saunders, J. Barrett, A. Kent, and D. Wallace, 227–263. Oxford: Oxford University Press.

Elementary Particles and Metaphysics

F.A. Muller

1 Polyphonic Prelude

LOTTERASI I inquire into reality, to find out what there is, how it behaves, how it is structured, how it all hangs together, independent of our cognitive and sensory capacities, independent of our longings and yearnings, independent of our actions and activities, independent of our very existence. Metaphysics is the philosophy of reality.

TANK I distinguish *noumenal reality* from the *phenomenal world*: the aforementioned is together with our faculties of understanding responsible for the last-mentioned. All scientific knowledge, which is synthetic a posteriori, is knowledge of the *phenomenal world*. Metaphysical knowledge, which is synthetic a priori, is about how we are able to pull it of, e.g. finding necessary conditions *we* have to meet in order *for us* to be able to know and to experience the world. Metaphysics is the philosophy of how we come to know "reality", an exploration of our faculties of understanding.

NARPAC I am with Lotterasi, Tank. Sort of. Questions about our faculties of understanding are questions about reality, and thereby are the proper subject-matter of *science*, notably cognitive psychology and cognitive neuroscience. Your kind of inquiry is arm-chair pseudo-science.

TANK But those branches of science you mention also have to presuppose the faculties they want to investigate, which spreads a smell of vicious circularity.

F.A. Muller (✉)
Faculty of Philosophy, Erasmus University Rotterdam, Burg. Oudlaan 50, 3062 PA Rotterdam, The Netherlands

Department of Physics, Institute for the History and Foundations of Science, Utrecht University, Budapestlaan 6, 3584 CD Utrecht, The Netherlands
e-mail: f.a.muller@fwb.eur.nl; f.a.muller@uu.nl

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_28,

NARPAC Not more viciously circular than a physician uses his lungs to breathe when he investigates lungs. There is nothing circular about it.

TANK What, then, is there left to do for metaphysics?

NARPAC Nothing. The story of metaphysics ends right here right now. Sorry Tank. The task of philosophy is clarification, in a broad sense of that term, of the means and products of science, and not to produce some kind of *philosophical* knowledge in addition to *scientific* knowledge.

NIQUE *Philosophy of science is philosophy enough!*

SLEWI Not by a long shot. Lotterasi is right, Tank is wrong and Narpac has lost it. Many issues in philosophy are metaphysical and science does not deal with them at all. Possible worlds! Necessitation! Chance! Endurantism! Perdurantism! Ontological dependence! Supervenience! Causal nexus! Diachronic identity! Synchronic identity! Comparative similarity! Universals! Tropes! Determinism! Bare particulars! Bundles! Mereology! Eternalism! Presentism! Intrinsic properties! Counterparts! External relations! Substantivalism! Relationism! Consciousness! Ontology!

NIQUE *Gavagai! Gavagai!*

NARPAC Please Will, crawl back to your web, scratch the hell out of your nerve endings, and only return when you have proved the consistency of your beloved New Foundations.

MUNTAP The current boom of analytic metaphysics, after it was presumed dead, is unintended consequence of Quine's "On what there is". Quine single-handedly made Ontology a respectable subject again. The phrase *ontological commitment* has captured the hearts and minds of the young. Science is admirable and then so is philosophy, metaphysics included, because science is continuous with science.

PEIRC *For Carnap*, metaphysical questions are a use-mention conflation, they can only be *about linguistic frameworks*; taken as questions about reality, they are *pseudo-questions*. Quine admittedly drove a stake through the heart of Carnap's "Empiricism, Semantics and Ontology". Yet the young are victim of a blatant misunderstanding of Quine: Carnap and Quine play in the same anti-metaphysical team! *For Quine*, metaphysical questions neither have pragmatic consequences nor are they in any way connected to scratchings at nerve endings, and therefore are empty and useless.

The reason why Analytic Metaphysics has flourished over the past decades and is flourishing in the present decade roughly is, according to Muntap and Peirc, that the philosophers engaged in analytic metaphysics have not understood Carnap and Quine properly.[1] We are witnessing here a misunderstanding on the grand scale. Analytic metaphysics is bogus.

[1]Lowe (2006), Westerhoff (2005), Cocchiarella (2007), and Price (2009), see further any current anthology and handbook on metaphysics, such as Gale (2002) and Kim et al. (2009).

This is not credulous. Part of an entire generation of philosophers *cannot read*? David Kellogg Lewis and Donald Williams *could not read*? David Armstrong, William Alston, David Chalmers, Jonathan Shaffer and Theodore Sider, say, *cannot read*? Of course they can. They have understood Carnap and Quine as well as anyone. But they are not *convinced*.

Why are they not convinced is a tall story, which they themselves can tell better than I ever could.

One reason for the return of metaphysics comes from philosophy of science, or so I speculate. We find it in the thriving realism debate. After the rise of realism during the 1960s and 1970s, the chains of sense data were shedded and the emperor of sensory experience no longer ruled over the minds of philosophers of science. A hostile attitude towards metaphysics was not in harmony with the spirit of realism. Further, interpretation problems of probability, modality, renewed debates about the nature of space, time and space-time in the wake of the general theory of relativity, of the nature of physical reality in the wake of the notorious measurement problem of quantum physics, all these subjects look pretty metaphysical and aroused interest in metaphysics.

In defence of analytic metaphysics, one can point out that philosophers engaged in metaphysical discourse have their own analytic linguistic framework, and they communicate as good, or as bad, with each other as other people do in their discourse. They propound theses, present analyses, propose explanations, expound arguments, perform thought experiments, pump up intuitions, elicit understanding, make intelligible, reason critically and constructively. What is the problem?

A critic of analytic metaphysics may now argue that, in contrast to what happened in scientific disciplines, philosophers engaged in analytic metaphysics are not making any *progress* in *their* linguistic framework. They have nothing to show for after decades of renewed metaphysical inquiry. Controversy and confusion dominate over consensus and clarity. The lack of progress should be sufficient to end it all. Why continue a discourse when there is no advancement and nothing but stagnation?

T.S. Kuhn showed that when meta-questions trouble the minds of scientists working in some *paradigm*, they have a *crisis* on their hands, which may be the death knell of the paradigm. Glossing over the fact that philosophy is not the same as (but may be continuous with) science, we may then see the currently heightened attention for *meta-metaphysical* issues in analytic metaphysics as an indication that it is in crisis.[2] The flight back to metaphysics, on the ruins of positivism, tired of the boredom of pragmatism, fed up with the servitude of science, may be short-lived, a last melancholic convulsion of a many centuries old and grand discourse. Or is, unlike in science, posing meta-questions part and parcel of philosophy and therefore *normal philosophy*?

Perhaps there is no crisis, or else such crises belong the essence of philosophy – and perhaps to the essence of philosophy only: that it is in a chronic state of

[2]Chalmers et al. (2009) and MacLaurin and Dyke (2012).

crisis. Just as every respectable rock band always is on the verge of breaking up, philosophy is always on the verge of revising its aims, methods and results. (Maybe this is why there isn't any progress in philosophy . . .).

The thesis that there is no progress in philosophy however is, generally speaking, false. Undisputably there is progress in logic and logic is a branch of philosophy. Undisputably there is progress in philosophy of physics and whenever some results in philosophy of physics can be classified as metaphysical, there is progress in metaphysics too. A fair amount of progress consists in lasting logical relationships between philosophical theses and important concepts.

The organisation of this prima facie somewhat wandering paper is as follows. In Sect. 2 we take a peek at "naturalised metaphysics" and call on philosophy of physics to play a pivotal role here. In Sect. 3, we inquire whether Falkenburg's tome on elementary particles is an instance of naturalised metaphysics. In Sect. 4, we report the announced steps forward in the philosophy of physics that qualify as metaphysical, and conclude this is indisputably an instance of naturalised metaphysics. In Sect. 5, we bring French's 2nd UnderDetermination Thesis (2UDT) on the stage and discuss its bearing on our reported steps forward.

2 Naturalised Metaphysics

Opening lines of the Preface of *Ever thing must go!* Ladyman and Ross (2008):

> This is a polemical book. One of its main contentions is that contemporary analytic metaphysics, a professional activity engaged in by some extremely intelligent and morally serious people, fails to qualify as part of the enlightened pursuit of objective truth, and should be discontinued. We think it is impossible to argue for a point like this without provoking some anger. Suggesting that a group of highly trained professionals have been wasting their talents – and, worse, sowing systematic confusion about the nature of the world, and how to find out about it – isn't something one can do in an entirely generous way.

They continue a little further:

> We recognize that we may be regarded as a bit rough on some other philosophers, but our targets are people with considerable influence rather than novitiates. We think the current degree of dominance of analytic metaphysics within philosophy is detrimental to the health of the subject, and make no apologies for trying to counter it.

Ladyman and Ross do not want to abolish metaphysics entirely, but rather want to halt stagnating discussions concerning particular issues. They adhere to "non-positivist verificationism" to pursue what they call *naturalised metaphysics*. They propound the two following sufficient conditions for when to take a metaphysical issue *not seriously*, so that it can be ignored justifiably (Ladyman and Ross (2008), p. 29):

- No philosopher should take seriously what is beyond current scientific means of investigation according to the current scientific community.

- No philosopher should take seriously what has no identifiable bearing on the relationship between at least two relatively specific hypotheses that are either accepted by, or motivated and regarded as in principle testable by, the relevant current scientific community.

For example, the existence of universals and of possible worlds, and the issue when objects have a mereological sum, do not belong to naturalised metaphysics, because there these issues have no identifiable bearing to any scientific hypothesis and is forever, it seems, beyond the scientific means of investigation. Discussion about these issues should end.

Ladyman and Ross (2008) advance two more detailed conditions, one for taking seriously and one for rejecting a Metaphysical Claim (MC):

Principle of Naturalistic Closure. If with the aid of some MC two currently accepted scientific propositions (of which at least one is taken from current physics) *explain more* (in the unification sense of Friedman and Kitcher) than separately and without MC, then MC should be taken seriously (p. 37).

The Primacy of Physics. If some MC about some special science conflicts with currently accepted fundamental physics, then reject MC (p. 44).

We are not going to analyse, defend, attack, illustrate or clarify these principles. We put them here on display as a serious attempt to draw the line between "good" and "bad" metaphysics, where naturalised metaphysics is identified as good. Is there any piece of metaphysics that is good in the Ladyman-Ross sense?

Ladyman and Ross themselves engage in the kind of metaphysics that is naturalised, by sketching a metaphysical view called *ontic structural realism*, with the concept of a *pattern* as its beating heart. This entire view then is (supposed to be) an MC that meets the Principle of Naturalistic Closure and is not in conflict with current fundamental physics (Primacy of Physics).

There is more naturalised metaphysics. Various debates conducted in philosophy of physics, e.g. substantivalism versus relationism, the reality problem (aka the measurement problem) in quantum physics, eternalism versus presentism with regard to time, the nature of elementary particles in quantum physics, get the stamp of approval by Ladyman and Ross. To the issue of elementary particles in quantum physics we want to turn next, in particular to B. Falkenburg's *Particle Metaphysics* (2007), which seems a place where to find naturalised metaphysics.

3 The Falkenburg Lists

The subtitle of Falkenburg (2007) reads: a Critical Account of Subatomic Reality. As it turns out, Falkenburg means 'critical' in some Kantian sense. Neo-Kantianism version n, for some unspecifiable natural number $n > 1$. A thorough inquiry into elementary particle physics leads Falkenburg to conclude that *there is no single particle concept in play*, but rather a somewhat startling variety of particle concepts,

which agree on some and differ on other aspects. Is, then, the concept of an elementary particle a family resemblance concept? We are going to take a brief look at these remarkable Falkenburg lists, which have drawn little attention from philosophers of physics and less from philosophers in metaphysics.

Classical Particles are:

MQ	carriers of mass and electric charge;
INDEP	independent of each other;
POINT	point-like in interactions;
CONS	obey conservation laws;
LOC	always localised;
DET	behaviour determined by the laws of classical mechanics;
TRAJ	moving on trajectories in phase space;
INDIV	spatio-temporally individuated (individuals);
BOUND	able to form bound composite physical systems.

INDIV presupposes space-time substantivalism, because for relationists in spatially symmetric configurations, particles fail to be spatio-temporal individuals, e.g. three particles following inertial worldlines that pass the corners of an equilateral triangle in every simultaneity hypersurface. Then these particles are not absolutely discernible and therefore not individuals (see further Sect. 4). Clearly INDIV is a metaphysical claim.

Quantum Particles are:

MQS	carriers of mass, electric charge and spin;
INDEP	independent of each other;
POINT	point-like in interactions;
CONS	obey conservation laws;
LOCD	localisable by a detector;
PROB	behaviour probabilistically determined by the laws of QM;
UNPQ	never jointly sharp in position and momentum;
PAULI	only distinguished by their quantum states (Exclusion Principle);
BOUND	able to form bound composite physical systems;
REPR	correspond to representations of space-time symmetry groups.

Falkenburg's assessment of condition PAULI is wrong: every particle in a composite system of similar particles is in *the same state* (partial trace) in symmetric as well as anti-symmetric states; quantum particles are not individuals, they are not absolutely discernible.

Light Particles are:

EN0	massless and chargeless carriers of energy ($E = h\nu$);
INDEP	independent of each other;
POINT	point-like in interactions;
CONS	obey conservation laws;
NLOC	non-local (have wave-vector: $\mathbf{k} = \mathbf{p}/h$);

LOCD	localisable by a detector;
PROB	behaviour of probability determined by classical radiation theory;
WAVE	in states that superpose and interfere;
BOSE	indistinguishable, BE-statistics.
DISC	discontinuous, i.e. they come in quanta.

Quantum particles also obey WAVE of the Light List and DISC of the Classical List, and classical particles obey DISC of the Light List. Light particles also have spin and they are not individuals, just as the quantum particles (INDIV).

Field Quanta (modes of excitation of quantum fields) are:

MESQ	collections of mass, energy, spin and charge;
INDEP	independent of each other;
POINT	point-like in interactions;
CONS	obey conservation laws;
NLOC	non-local;
LOCD	localisable by a detector;
PROB	behaviour probabilities determined by quantum field eqs.;
WAVE	in states that superpose and interfere (Fock space);
COMM	subject to anti-comm. or comm. rules;
INDIST	only numerically distinguishable;
DISC	discontinuous, i.e. they come in quanta;
BOUND	able to form bound composite physical systems.

Field quanta are indiscernibles. Also from quantum field theory come *virtual particles*, which are:

MESQ	collections of mass, energy, spin and charge;
DEP	dependent on interacting quantum particles, no independent existence;
NCONS	violate energy conservation laws, e.g. *off mass shell*;
NLOCD	not localisable by a detector;
WAVE	in states that superpose and interfere;
INDIST	indistinguishable;
DISC	discontinuous, i.e. they come in quanta;
PATH	follow path-integrals;
PERT	occur in perturbation expansions of scattering matrix, in Feynman diagrams representing propagators.

These virtual particles are even more elusive than the field quanta. A question like "How many virtual particles are there at time t?" makes no sense.

Quasi-particles (phonons, plasmons, ecxitons, . . .) are:

EMESQ	collections of *effective* mass, energy, spin and charge;
INDEP	independent of each other;
CONS	obeys conservation laws;

WAVE	in states that superpose and interfere;
COMM	subject to anti-comm. or comm. rules;
COLL	collective excitations of a ground-state;
UNPQ	unsharp joint momentum and location;
INDIST	indistinguishable;
DISC	discontinuous, i.e. they come in quanta.

What to conclude from these lists? *First* of all, we can list what classical particles, quantum particles, light particles and field quanta have in common. Let's call these **Particles**:

PROP	they have some intrinsic properties;
INDEP	independent of each other;
POINT	point-like in interactions;
CONS	obey conservation laws;
LOCD	localisable by a detector;
DISC	discontinuous, i.e. they come in quanta (of matter or radiation).

Thus the concept expressed by this short list remains stable through the transition from classical to quantum mechanics (QM) and from QM to quantum field theory (QFT). So Particles have survived the mentioned transitions and this continuity points into the direction of entity realism.

Secondly, virtual particles and quasi-particles arguably "are not really particles". It seems a stretch to maintain that their lists of conditions are sufficient to speak of "particles". They are definitely not Particles because, for one thing, they violate LOCD. Whether being localised in a "small" finite region, or even in a point, at any time (in every frame of reference), which is LOC, is essential for being a particle is an important and controversial subject. Malament (1996) and Halvorson and Clifton (2002) have insisted on LOC as essential. Then there are particles neither in QM nor in QFT, for they both violate LOC. But according to Falkenburg's lists, there are particles both in QM and in QFT, because quantum particles and field quanta, respectively, *are* Particles. Certainly in QM position has lost its central role as a physical magnitude when compared to classical mechanics (CM): whereas in CM the worldlines of particles in space-time are the solutions of the dynamical eq. of CM, which contain all information about the particles that CM provides (save their mass in case of inertial worldlines), there are no worldlines in QM and the solutions of the dynamical eq. of CM are curves in Hilbert-space. Followers of taking LOC as essential for particles clash with modern physics, which has a well-delineated particle concept. Dilemma: must we, as philosophers of science and of physics, (i) charge modern physicists with conceptual confusion and ontological delusion because of their persistent talk and detection of particles whereas they have no particle concept and thus are babbling incoherently when they utter the word "particle", or (ii) conclude that *there are* particles in QM and QFT, as in CM, just as there are bears in America, Asia and Europe, but that the particles in QM, QFT and CM differ in kind, just as the cinnamon bears of Colorado, the Tibetan blue bears and the brown bears of the Pyrenees differ in subspecies. Falkenburg has chosen horn (ii), as did Saunders (1994) and Wallace (2006) before her.

Thirdly, we can ask what distinguishes quantum particles from classical particles, and we obtain a rather precise answer: classical particles obey conditions LOC, DET, TRAJ, INDIV and MBS, and do not obey: PROB, LOCD, WAVE, INDIST, FDS, BES and REPR, whereas for quantum particles we have exactly the opposite. So although both classical and quantum particles are Particles, they also differ significantly.

Some of Falkenburg's own claims drawn from here lists are worth reporting:

- From classical to quantum particle is an intensional change, not an extensional one.
- From classical particle to light quantum is an intensional change as well as an extensional one.
- Infinitely many virtual particles together have causal power.
- Quasi-particles are not as fictitious as virtual particles they come on their own as quantised appearances.

In the final Chapter, Kant of Königsbergen mounts the stage after a thorough make-over from Bohr of Copenhagen: subatomic reality consists of phenomenal quantum structures. Phenomenal subatomic reality is thoroughly relational: (a) observation of phenomena depends on experimental arrangements; all quantum phenomena occur in a classical world; (b) is "defined" relative to classical concepts; (c) energy-dependent; and more.

Thus Falkenburg clearly brings here the results of her conceptual analyses from elementary particle physics into contact with Kantian metaphysics, and also with the realism debate in the philosophy of science, but not with contemporary analytic metaphysics. The Falkenburg lists may very well bear on issues in contemporary analytic metaphysics, such as the issues of ontological dependence, ontological grounding, fundamentality, and quantifier domains. There is no trace of any serious engagement in Falkenburg's tome (2007) with these issues. French (2009) pointed out there is also no trace of engagement with recent discussions in the philosophy of physics about whether in QFT a particle concept can be maintained in cases of interactions – Falkenburg only seems to have considered asymptotic free scattering states, where one has (approximately) free particles. Of course not the end of the world this is. Falkenburg did not enter the discourse of analytic metaphysics with her lists and conclusions, although they carry the mark of naturalised metaphysics.

Looked upon Falkenburg's inquiries from analytic metaphysics, one must ascertain that her tome is pretty much ignored.

Thus it seems that analytic-metaphysical discourse proceeds in splendid or dangerous isolation from philosophical discourse of physics. This is awkward. Physics is the branch of science that is *most general* (Dummett) and has *complete coverage* (Quine) and provides us with knowledge of the most general kind available in science. Which topics in analytic metaphysics could profit from Falkbenburg's results?

Topic 1. Objecthood. The concept of an object is a topic of metaphysical attention. Do the various kinds of particles on the Falkenburg Lists qualify as objects? Do these Lists necessitate a re-conceptualisation of objects?[3]

Topic 2. Independence. Is there a relation between the concept of ontological independence of metaphysics and item INDEP on Falkenberg's lists?

Topic 3. Leibniz's Principle of the Identity of Indiscernibles. To this topic the next Section is devoted.

4 The Rise of Relationals

Hermann Weyl (1928, iv.c. Sect. 9; 1949, p. 247) noticed that the permutation symmetry of quantum mechanics (QM) seems to make QM inconsistent with Leibniz's Principle of the Identity of Indiscernibles (PIdIn): necessarily, if the things are indiscernible, then they are identical. The converse, aka Leibniz's Law, is a logical necessity. If QM, a well-confirmed physical theory, is incompatible with PIdIn, then PIdIn is neither a nomic nor a metaphysical necessity, and should go. We have a clash here between science and metaphysics. Weyl only advanced this for bosons; he thought that fermions comply to PIdIn because of Pauli's Exclusion Principle ("no two electrons are in the same state"). Henry Margenau (1944, p. 202) put Weyl aright: the states of every particle in systems of similar particles, whether they be fermions or bosons, are identical partial traces due to the Symmetrisation Postulate. In all atoms and molecules electrons *are* in the same state. (Pauli's Exclusion Principle really states that no $\otimes$-factor in a term of a permutation-symmetric pure state occurs more than once.) A remarkable oversight of Weyl. Thus the *Incompatibility Thesis* (of QM and PIdIn) became firmly entrenched in the philosophy of physics. Then B.C. van Fraassen (1984) discovered that the standard Property Postulate of QM was needed in order to have a decent deductive argument for the Incompatibility Thesis, which Postulate, when applied to particles, reads as follows: a particle has a property represented by $\langle A, a\rangle$, where A is some operator representing a physical magnitude, and where $a \in \mathbb{R}$ is a value from the spectrum of A, iff the state of the particle is an eigenstate of A having eigenvalue a. Notwithstanding this discovery, the Incompatibility Thesis was hammered home by an army of philosophers (of physics).[4]

[3] See Lowe (2006).

[4] Schrödinger (1996). Cortes (1976), who brandished PIdIn "a false principle", Barnette (1978), Ginsberg (1981), French and Redhead (1988), and Giuntini and Mittelstaedt (1989), who argued that although demonstrably valid in classical logic, in quantum logic the validity of PIdIn cannot be established, French (1989), who assured us that PIdIn "is not contingently true either", French (2009), French and Rickles (2003), Redhead and Teller (1992), Butterfield (1993), Teller (1998), Castellani and Mittelstaedt (1998), and Huggett (2003); refinements and elaborations have been appearing ever since, lately delving into the metaphysics of "object" and "individual", e.g. French and Redhead (1988) and French and Krause (2008).

The tide began to turn when S.W. Saunders (2006), and Muller and Saunders (2008) proposed to extend the sufficient condition for identity in PIdIn from indiscernibility by *properties* to also include indiscernibility by *relations*, following Quine (1976), who had earlier inquired into the many ways of discernibility. Saunders further propounded that two fermions in the permutation-symmetric singlet state are differently related to themselves than to each other by the relation "has opposite spin to". This was later extended and proved for all fermionic and bosonic composite systems. They do meet the sufficient condition for identity and the lamentable conclusion that we have only one particle can no longer be deduced.

Some terminology now. Call an object *absolutely discernible*, or an *individual*, iff it has some property that all other particles lack; that property, then, is "its individuality". Call an object *relationally discernible* iff it is *not* related to all objects in the same way, itself included; call an object a *relational* iff it is relationally discernible but not absolutely discernible. Call an object *indiscernible* iff it is neither absolutely nor relationally discernible. Finally, call two objects *witness-discernible* iff they are differently related to some third object, the witness, and to all other objects that are not absolutely discernible from the witness.[5] The Principle of the Identity of *Absolute* Indiscernibles (PIdAIn) reads: necessarily, objects that are absolutely indiscernible are identical. PIdIn becomes: necessarily, if objects are absolutely and relationally indiscernible, then they are identical. Leibniz was the first to discuss PIdAIn elaborately and to apply it to "substances"; in several places, Leibniz defends (as we would put it today) a reduction of relations to properties, which makes mentioning relations in discernibility otiose.[6]

Against a background of classical logic, Quine (1976) demonstrated there are only *two* kinds of relational discernibility: *relative discernibility* (the discerning relation is anti-symmetric) and *weak discernibility* (the discerning relation is symmetric and irreflexive). Ladyman et al. (2012) proved that: absolute discernibility implies relative discernibility implies weak discernibility implies distinctness, but all converse implications fail. Linnebo and Muller (2012) demonstrated that witness-discernibility collapses onto absolute discernibility and therefore is not a novel category of discernibility. Muller and Seevinck (2009) showed that for both bosons and fermions one can define relations, in the language of QM, that are physically significant and permutation-invariant, and that discern all particles weakly. Thus rises the novel metaphysical category of relationals.

These results surely count as progress in philosophy of physics and in metaphysics; we now know something that we did not know before: that there are a number of demonstrably different kinds of discernibility, how they hang together logically, and that elementary particles are *relationals*, so that relations turn out to

[5] Dieks and Versteegh (2008) and Ladyman and Bigaj (2010).

[6] See Russell (1937, pp. 13–15) and Ishiguro (1990, pp. 118–122, 130–142) for Leibniz's struggle with relations.

populate the universe, and that PIdIn is *not* in conflict with modern physics – in QM, PIdIn becomes a *theorem* when expressed in the language of QM. These steps forward also count as naturalised metaphysics, we submit.

This is not the end of it. The 2nd UnderDetermination Thesis threatens to be a spoiler.

5 The Second UnderDetermination Thesis

The Duhem-Quine 1st UnderDetermination Thesis (1UDT) roughly says that all our observations do not or even cannot determine which scientific theory is true. With every set of experimental results, data, a plethora of distinct theories is compatible. The 2nd UnderDetermination Thesis (2UDT), expounded by French, says that theories do not or even cannot determine a metaphysical picture of reality, which is a "picture" that answers questions that are addressed in metaphysics, rather than leaves them open. With every scientific theory, a plethora of distinct metaphysical views is compatible. The various interpretations of QM illustrate 2UDT rather shiningly. French argued that also what QM says about elementary particles is compatible with different Metaphysical Views about the nature of elementary particles.

(MV1) Elementary particles are indiscernibles – PIdIn stands refuted, change in background mathematics or logic becomes necessary, move from set-theory to qset-theory or from classical or intuitionistic logic to some deviant logic, like "Schrödinger logic".

(MV2) Elementary particles are relationals – PIdIn saved, nothing changes.

(MV3) Elementary particles are individuals – PIdAIn saved, cloth QM with Caulton's recent basis-dependent ontology or reject QM altogether and adopt Bohmian Mechanics.

Now what? When propounders of these different views do not reach agreement about which view is, all things considered, the best view, it is difficult to maintain that metaphysical progress has been achieved about microphysical reality. Rather we are stuck in controversy. Let's see what we say can about this.

We begin by noticing that whereas 1UDT can, besides by means of examples from the history of science and contemporary science, also be based on rigorous logical arguments, 2UDT can only be based on examples, and is therefore predominantly a *descriptive* thesis. Next we notice that in the case under consideration here, MV2 is grounded in theorems that have been deduced from a few postulates of QM. Therefore everyone who accepts: QM, the relations used in the mentioned proofs, and deductive logic is committed to these theorems, and thereby to MV2. The ways for proponents of MV1 and MV3 then are: (i) to reject QM (MV3), or (ii) to reject the weakly discerning relations for some reason, or (iii) point to a flaw in the proofs, or (iv) to adopt some deviant mathematics or logic that will block the deduction of these theorems (MV1) but keeps other deduced results. Some have criticised the proofs of MV2 for being circular; but these shots turned out

to be fired with blanks.[7] So much for option (iii). Option (iv) is drastic if not an overreaction, for which there are no independent arguments. Option (ii) can be taken as a stark rejection of relations to discern and to defend that only properties can discern. Option (i) is not on the table for a naturalist metaphysician. Proponents of naturalised metaphysics – the kind we discussed in Sect. 2 – glide naturally to MV2. No additional metaphysical shoring up is needed. Perhaps naturalised metaphysical views do follow naturally from scientific theories and 2UDT fails for them.

But a naturalist philosopher may also pass over in silence and prefer to remain neutral on an issue whenever science does not determine it, even if it does follow naturally from it. If 2UDT is true, then we should not waste time pondering *which* metaphysical view to choose altogether, because to engage in such debates then might be an act of betrayal of naturalism.

Another option entirely is to *break* this particular underdetermination by making the familiar Ramseyan move by rejecting a common presupposition of all metaphysical views MV1–MV3, which is that there are particles. Dieks and Lubberdink (2011) argued that there are no particles in QM but that in wave-functions with peaked position-probability distributions we can legitimately speak of "classical particles". Whether or not there are "classical particles" then is a contingent matter, depending on the wave-function of the composite system. In cases where there are "classical particles", the Ramseyan move is of no avail. The only currently known way to make the Ramseyan move is to erase particles from our fundamental ontology entirely. Enter *ontic structural realism*: there are no particles at the fundamental level of physical reality, there are only structures and object-like features of structures.[8] What seems the fundamental substance of physical reality are quantum fields.[9] Quantum fields are structures. They have, perhaps, object-like features: the *field quanta* as characterised by Falkenburg (Sect. 3).

6 Recapitulation

We began, in our Polyphonic Prelude, to rehearse a few very well known views on the nature of metaphysics, leading to the currently flourishing field of analytic metaphysics. We wondered whether it is in a state of crisis, possibly indicated by the sudden attention to meta-metaphysics. Whether there is progress and even whether progress is possible was an issue. Somehow connected seems the fierce criticism of analytic metaphysics by Ladyman and Ross, which we considered next. This led us a kind of metaphysics, naturalised metaphysics, that does not fall prey to their acerbic criticism. Philosophers of physics then seemed to engage – regularly at least – in naturalised metaphysics. We then looked at Falkenburg's inquiry into the nature of elementary particles. While the main value of her tome predominantly

[7] Hawley (2009), Muller and Seevinck (2009) and Muller (2014).

[8] Ladyman (1998), French and Ladyman (2003), and French and Krause (2008).

[9] See Kuhlmann et al. (2002) for metaphysical struggles with quantum field theory.

lies in her conceptual analyses of the various particle concepts which are present in modern physics – indisputably an instance of progress –, several possibilities to establish tighter connexions between philosophy of physics and analytic metaphysics remained wide open, e.g. ontological dependence and the issue of Leibniz's Principle of the Identity of Indiscernibles. We then went on to report more progress where philosophy of physics and metaphysics meet and went on to a threat to some of steps forward in the form of 2UDT. We avoided the threat by a Ramseyan move and ended at the gates of ontic structural realism. Home sweet home.

Acknowledgements Thanks to Dennis Dieks.

References

Barnette, R.L. 1978. Does quantum mechanics disprove the principle of the identity of indiscernibles? *Philosophy of Science* 45: 466–470.

Brading, K., and E. Castellani (eds.). 2003. *Symmetries in physics: New reflections*. Cambridge: Cambridge University Press.

Butterfield, J.N. 1993. Interpretation and identity in quantum theory. *Studies in History and Philosophy of Science* 24: 443–476.

Castellani, E., and P. Mittelstaedt. 1998. Leibniz's principle, physics and the language of physics. *Foundations of Physics* 30: 1587–1604.

Chalmers, D., D. Manley, and W. Wasserman (eds.). 2009. *Metametaphysics: New essays on the foundations of ontology*. Oxford: Clarendon Press.

Cocchiarella, N.B. 2007. *Formal ontology and conceptual realism*. Dordrecht: Springer.

Cortes, A. 1976. Leibniz's principle of the identity of indiscernibles: A false principle. *Philosophy of Science* 43: 491–505.

Dieks, D., and A. Lubberdink. 2011. How classical particles emerge from the quantum world. *Foundations of Physics* 41: 1051–1064.

Dieks, D., and M.A.M. Versteegh. 2008. Identical quantum particles and weak discernibility. *Foundations of Physics* 38: 923–934.

Falkenburg, B. 2007. *Particle metaphysics. A critical account of subatomic reality*. Berlin: Springer.

French, S. 1989. Why the principle of the identity of indiscernibles is not contingently true either. *Synthese* 78: 141–166.

French, S. 2009. Review of Falkenburg (2006). *Studies in the History and Philosophy of Modern Physics* 40: 194–195.

French, S., and D. Krause. 2008. *Identity in physics: A historical, philosophical and formal analysis*. Oxford: Clarendon Press.

French, S., and D. Rickles. 2003. Understanding permutation symmetry. In *Symmetries in physics: New reflections*, ed. K. Brading and E. Castellani, 212–238. Cambridge: Cambridge University Press.

French, S., and J. Ladyman. 2003. Remodelling structural realism: Quantum physics and the metaphysics of structure. *Synthese* 136: 31–56.

French, S., and M.L.G. Redhead. 1988. Quantum physics and the identity of indiscernibles. *The British Journal for the Philosophy of Science* 39: 233–246.

Gale, R. (ed.). 2002. *The Blackwell guide to metaphysics*. Oxford: Blackwell.

Ginsberg, A. 1981. Quantum theory and the identity of indiscernibles revisited. *Philosophy of Science* 48: 487–491.

Giuntini, R., and P. Mittelstaedt. 1989. The Leibniz principle in quantum logic. *International Journal of Theoretical Physics* 28: 159–168.

Halvorson, H., and R.K. Clifton. 2002. No place for particles in relativistic quantum theories? *Philosophy of Science* 69: 1–28.

Hawley, K. 2009. Identity and indiscernibility. *Mind* 118: 102–119.
Huggett, N. 2003. Quarticles and the identity of indiscernibles. In *Symmetries in physics: New reflections*, ed. K. Brading and E. Castellani, 212–238. Cambridge: Cambridge University Press.
Ishiguro, H. 1990. *Leibniz's philosophy of logic and language*, 2nd ed. Cambridge: Cambridge University Press.
Kim, J., E. Sosa, and G.S. Rosenkrantz. 2009. *A companion to metaphysics*. Oxford: Blackwell.
Kuhlmann, M., et al. 2002. *Ontological aspects of quantum field theory*. River Edge: World Scientific.
Ladyman, J. 1998. What is structural realism? *Studies in the History and Philosophy of Science* 29: 409–424.
Ladyman, J., and D. Ross. 2008. *Every thing must go. Metaphysics naturalised*. Oxford: Oxford University Press.
Ladyman, J., and T. Bigaj. 2010. The principle of the identity of indiscernibles and quantum mechanics. *Philosophy of Science* 77: 117–136.
Ladyman, J., Ø. Linnebo, and R. Pettigrew. 2012. Identity and discernibility in philosophy and logic. *Review of Symbolic Logic* 5(1): 162–186.
Linnebo, Ø., and F.A. Muller. 2012. On witness-discernibility of elementary particles. *Erkenntnis*. doi:10.1007/s10670-012-9385-4.
Lowe, J. 2006. *The four-category ontology. A metaphysical foundation for natural science*. Oxford: Clarendon Press.
MacLaurin, J., and H. Dyke. 2012. What is metaphysics for? *Australasian Journal of Philosophy* 90(2): 291–306.
Malament, D.B. 1996. In defence of dogma: Why there cannot be a relativistic quantum mechanics of particles. In *Perspectives on quantum reality*, ed. R.K. Clifton, 1–10. Dordrecht: Kluwer.
Margenau, H. 1944. The exclusion principle and its philosophical importance. *Philosophy of Science* 11: 187–208.
Muller, F.A. 2014. The rise of relationals. *Mind*, to appear.
Muller, F.A., and S.W. Saunders. 2008. Discerning fermions. *The British Journal for the Philosophy of Science* 59: 499–548.
Muller, F.A., and M.P. Seevinck. 2009. Discerning elementary particles. *Philosophy of Science* 76: 179–200.
Price, H. 2009. Metaphysics after Carnap: The ghost who walks? In *Metaphysics*, ed. D. Chalmers, D. Manley, and W. Wasserman, 320–346. Oxford: Oxford University Press.
Quine, W.V.O. 1976. Grades of discriminability. *Journal of Philosophy* 73: 113–116.
Redhead, M.L.G., and P. Teller. 1992. Quantum physics and the identity of indiscernibles. *British Journal for the Philosophy of Science* 43: 201–218.
Russell, B. 1937. *A critical exposition of the philosophy of Leibniz*, 2nd ed. London: George Allen & Unwin Ltd.
Saunders, S.W. 1994. A dissolution of the problem of locality. In *Proceedings of the philosophy of science association*, East Lansing, vol. 2, 88–98.
Saunders, S. 2006. Are quantum particles objects? *Analysis* 66: 52–63.
Schrödinger, E. 1996. *Nature and the greeks and science and humanism*. Cambridge: Cambridge University Press.
Teller, P. 1998. Quantum mechanics and haecceities. In *Interpreting bodies. Classical and quantum objects in modern physics*, ed. E. Castellani, 114–141. Princeton: Princeton University Press.
van Fraassen, B.C. 1984. Indistinguishable particles. In *Interpreting bodies. Classical and quantum objects in modern physics*, ed. E. Castellani, 73–92. Princeton: Princeton University Press.
Wallace, D. 2006. In defence of Naïveté: The conceptual status of Lagrangian quantum field theory. *Synthese* 151: 53–80.
Westerhoff, J. 2005. *Ontological categories*. Oxford: Clarendon Press.
Weyl, H. 1928. *Gruppentheorie und Quantenmechanik*. Leipzig: Hirzel.
Weyl, H. 1949. *Philosophy of mathematics and natural science*. Princeton: Princeton University Press.

Assessing the Status of the Common Cause Principle

Miklós Rédei

1 Introductory Comments

Since Kant's awakening from his dogmatic slumber it has been known that general metaphysical claims cannot be verified. Nor can they be conclusively falsified empirically; especially not, as Popper has taught us, if they are pure existential claims: one cannot be sure that a certain entity does *not* exist or that a certain state of affairs is absent from the world – simply because it is impossible to check empirically the whole universe, past, present and future, to make sure the entity in question does not exist or that a particular condition never obtains.

Given this well-known lesson from history of philosophy one should be very careful when it comes to assessing the metaphysical claim about the causal structure of the world known as the Common Cause Principle. The Common Cause Principle states that if two events A and B are probabilistically correlated, then either there is a direct causal link between A and B that is responsible for the correlation, or there exists a third event C, a common cause, that brings about the correlation. The Common Cause Principle is clearly a pure existential claim about causal connections lurking behind correlations and if it could be shown to be false, then one would have falsified not only the Common Cause Principle but even the meta-principle that metaphysical principles stating existence are non-falsifiable. This would be interesting; however, the meta-principle pronouncing the non-falsifiability of general existence claims does not seem to have been falsified by the Common Cause Principle: the aim of this paper is to argue that assessing the status of the Common Cause Principle is a very subtle problem, and that at present we do not have strictly empirical evidence that it does not hold.

M. Rédei (✉)
Department of Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London WC2A 2AE, UK
e-mail: m.redei@lse.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_29,

The complete argument that does justice to the complexity of the issue requires a full book (Hofer-Szabó et al. 2013). In this short paper just the main ideas and the most important concepts and claims are presented in a sketchy manner, typically without giving precise definitions of all the concepts involved. Specifically, standard mathematical notions from measure theory and operator algebra theory are used without explanation. (The Appendix in Hofer-Szabó et al. (2013) collects the most crucial mathematical facts involved.)

2 The Common Cause Principle

Given a classical probability measure space $(X, \mathcal{S}, p)$ with the set X of elementary events, Boolean algebra $\mathcal{S}$ of some subset of X and probability measure p, the events $A, B \in \mathcal{S}$ are called positively correlated if

$$p(A \cap B) > p(A)p(B) \tag{1}$$

Reichenbach's Common Cause Principle: If A, B are correlated events then either the events A and B stand in a direct causal relation responsible for the correlation, or, if A and B are causally independent, $R_{ind}(A, B)$, then there exists a third event C causally affecting both A and B, and it is this third event, the so-called *(Reichenbachian) common cause*, which brings about the correlation by being related to A and B in a specific way spelled out in the following definition:

Definition 1. C is a *common cause* of the correlation (1) if the following (independent) conditions hold:

$$p(A \cap B|C) = p(A|C)p(B|C) \tag{2}$$

$$p(A \cap B|C^{\perp}) = p(A|C^{\perp})p(B|C^{\perp}) \tag{3}$$

$$p(A|C) > p(A|C^{\perp}) \tag{4}$$

$$p(B|C) > p(B|C^{\perp}) \tag{5}$$

where

$$p(X|Y) \doteq \frac{p(X \cap Y)}{p(Y)}$$

denotes the conditional probability of X on condition Y, and it is assumed that none of the probabilities $p(X)$ $(X = A, B, C, C^{\perp})$ is equal to zero.

The above notion of common cause is due to Reichenbach (1956), the Common Cause Principle was articulated especially by Salmon (see e.g. Salmon 1984) and it has been discussed extensively in the literature. The bibliography in Hofer-Szabó

et al. (2013) contains an extensive list of papers on the topic, papers published in the last 10 years include Butterfield (2007), Cartwright (2007), Henson (2005), Hofer-Szabó (2008), Hofer-Szabó (2011), Hofer-Szabó et al. (1999), Hofer-Szabó et al. (2002), Hoover (2003), Mazzola (2012a), Mazzola (2012b), Portmann and Wüthrich (2007), Rédei and San Pedro (2012), Sober (2008), Sober and Steel (2012), San Pedro (2008), Wrónski (2010), Wrónski and Marczyk (2010), Wrónski and Marczyk (2013) and Wüthrich (2004).

3 How to Assess the Status of the Common Cause Principle?

Unless one takes the very problematic position that the truth of non-analytic statements can be decided by a priori reasoning without having a look at the world, the way to assess the status of the Common Cause Principle is to turn to our best descriptions of the world, namely to our confirmed scientific theories, to see if they do comply with the Common Cause Principle by being causally complete in the sense of providing a causal explanation of all the correlations they predict – either in terms of causal connections between the correlated entities or by displaying a common cause of the correlations. Doing so one can encounter the following cases in principle:

1. *Our probabilistic theories are causally complete*.
   This is confirming evidence for the Common Cause Principle – but not *proof* of its truth because our theories can be incomplete: not describing all the correlations that exist, and some of those they do not describe may not have a causal explanation.
2. *Our probabilistic theories are causally incomplete*.
   There are two subcases of this latter situation:
   a. A theory is causally incomplete but causally completable.
      These theories do not provide disconfirming evidence for the Common Cause Principle.
   b. A theory is causally incomplete and causally incompletable.
      Only such theories provide disconfirming evidence for the Common Cause Principle.

By causal completion of a probabilistic theory T that models the world in terms of a classical probability space $(X, \mathcal{S}, p)$ is meant a causally complete theory T' that describes the probabilistic aspect of the world by another probability measure space $(X', \mathcal{S}', p')$ that is an extension of the probability space $(X, \mathcal{S}, p)$ in the sense that there exists a Boolean algebra homomorphism $h: \mathcal{S} \to \mathcal{S}'$ that preserves the measure: $p'(h(A)) = p(A)$ for all $A \in \mathcal{S}$.

The essential message of this paper is that probabilistic theories are either causally complete, or, if they are not, then they *are* causally completable; furthermore, if we require reasonable, well-motivated conditions on the common cause *in*

addition to the four conditions in the definition of common cause, which obviously make causal completability of probabilistic theories more difficult, then it is an open problem whether good candidates for causally incomplete theories (quantum theory) are in fact causally incomplete, and if so, whether they are causally incompletable.

In view of the above it is therefore of interest to find out if probabilistic theories are causally complete. A particularly strong form of causal completeness is *common cause completeness*: a probability space $(X, \mathcal{S}, p)$ is common cause complete by definition if it contains a common cause of every correlation it predicts and it is called *common cause completable* with respect to a pair (A, B) of correlated events A, B in it if there exists an extension $(X', \mathcal{S}', p')$ of $(X, \mathcal{S}, p)$ such that $(X', \mathcal{S}', p')$ contains a common cause of the correlation between A and B. Call a probability space *strongly* common cause completable with respect to a pair (A, B) of correlated events A, B in it if, given any *type* of a common cause, there exists an extension $(X', \mathcal{S}', p')$ of $(X, \mathcal{S}, p)$ such that $(X', \mathcal{S}', p')$ contains a common cause of the given type of the correlation between A and B. "Type" of the common cause refers here to some additional probabilistic constraints one can in principle impose on the common cause (see Hofer-Szabó et al. 1999 for details).

It is not difficult to see that probability spaces with a finite number of elementary events are typically *not* common cause complete, not even causally complete with respect a causal independence relation R_{ind} satisfying certain plausible conditions (Gyenis and Rédei 2004, 2011a,b). "Large" probability spaces (probability spaces with an uncountably infinite Boolean algebra) are however common cause complete:

Proposition 1 (Gyenis and Rédei 2011).

1. *If a classical probability space is purely nonatomic as a measure space, then it is common cause complete.*
2. *Every classical probability measure space has an extension that is purely nonatomic; hence:*
3. *Every classical probability measure space is common cause completable with respect to any set of correlated events.*

Note that (3) in Proposition 1 does *not* say that every probability space is *strongly* common cause completable; whether this is true, is not known. The conjecture is that measure theoretically purely nonatomic probability spaces are *strongly* common cause complete: containing a common cause of *every* admissible type of *every* correlation they predict – by Proposition 1 this would entail that any probability space is strongly common cause completable with respect to every correlation. We only have a weaker result so far[1]:

[1] After this paper was completed, the author was informed that the conjecture has been proved by Marczyk and Wrónski (2013). See also their review paper in this volume: M. Marczyk and L. Wrónski: "A Note on Strong Causal Closedness and Completability of Classical Probability Spaces".

Proposition 2 (Hofer-Szabó et al. 1999). *Every classical probability measure space is strongly common cause completable with respect to any finite set of pairs of correlated events.*

The significance of the above two propositions is that they entail that Reichenbach's Common Cause Principle can always be defended against attempts of falsification by claiming that a correlation predicted by a probability theory is due to a possibly hidden common cause – hidden in the sense of not being part of the original theory predicting the correlation.

Thus one has to be epistemically modest when assessing the truth status of the Common Cause Principle: one only can aim at finding out whether scientific theories provide confirming or disconfirming evidence for the Principle, and the propositions above also entail how one can in principle display disconfirming evidence for the Common Cause Principle:

1. Impose extra conditions on the common cause that are not part of Reichenbach's definition.
2. Argue that the additional conditions are justified.
3. Display a theory T that predicts empirically testable correlations common causes of which should satisfy the extra conditions.
4. Show that theory T is causally incomplete.
5. Prove that theory T is not extendable into a richer theory that contains common causes satisfying the extra conditions.

Good candidates for the above are quantum correlations, common causes of which should be *local* in some sense; it will be argued however in the next section that (4) and (5) have *not* been shown yet for quantum theory. In case of quantum correlations one has to be careful however what one means by a common cause: as we have seen, Reichenbach's definition of common cause was formulated originally within the framework of classical, Kolmogorovian probability measure spaces, and quantum theory can be viewed as non-classical probability theory.

A general non-classical probability theory is a pair $(\mathcal{L}, \phi)$ where $\mathcal{L}$ is an orthocomplemented, orthomodular but non-distributive lattice (with respect to lattice operations $\wedge, \vee, \perp$) replacing the Boolean algebra of events and ϕ is a countably additive map (generalized probability) measure from $\mathcal{L}$ into the interval $[0, 1]$. Special cases of generalized probability spaces are the quantum probability spaces $(\mathcal{P}(\mathcal{N}), \phi)$ with $\mathcal{P}(\mathcal{N})$ being the lattice of projections of a von Neumann algebra $\mathcal{N}$ and ϕ being a normal state on $\mathcal{N}$. Standard Hilbert space quantum theory is an even more special case where $\mathcal{N} = \mathcal{B}(\mathcal{H})$ is the set of all bounded operators on a Hilbert space $\mathcal{H}$.

Given a general probability measure space $(\mathcal{L}, \phi)$, two *compatible* elements $A, B \in \mathcal{L}$ are called correlated in ϕ if

$$\phi(A \wedge B) - \phi(A)\phi(B) > 0 \tag{6}$$

One can define then the common cause in $\mathcal{L}$ of the correlation (6) by reformulating the conditions (2)–(5) without modification if one requires the common cause to commute with both A and B:

Definition 2. $C \in \mathcal{L}$ is called a common cause of the correlation (6) if C is compatible with both A and B and the following conditions hold.

$$\phi(A \wedge B|C) = \phi(A|C)\phi(B|C) \tag{7}$$

$$\phi(A \wedge B|C^{\perp}) = \phi(A|C^{\perp})\phi(B|C^{\perp}) \tag{8}$$

$$\phi(A|C) > \phi(A|C^{\perp}) \tag{9}$$

$$\phi(B|C) > \phi(B|C^{\perp}) \tag{10}$$

where

$$\phi(X|Y) \doteq \frac{\phi(X \wedge Y)}{\phi(Y)} \tag{11}$$

Notions such as type of the common cause, (strong) causal and common cause completeness, and (strong) causal and (strong) common cause completeness can now be defined in non-classical probability measure spaces in complete analogy with the classical definitions, and one is then led to a number of problems, some of which are still open. It is important however that the non-commutative analogue of Proposition 1 holds:

Proposition 3 (Kitajima 2008; Gyenis and Rédei 2013).

1. *If a non-classical probability space is purely nonatomic as a measure space, then it is common cause complete.*
2. *If a $\mathcal{L}$ is a σ-complete non-atomic lattice and ϕ is a faithful general probability measure on $\mathcal{L}$, then the probability space $(\mathcal{L}, \phi)$ is measure theoretically purely non-atomic.*

Note that a probability space (both classical and generalized) *can* be common cause closed and *not* purely non-atomic as a measure space, measure theoretic non-atomicity is just sufficient but not necessary for the theory to be common cause complete: if the space has at most one measure theoretic atom, then, and only then it is common cause complete (Gyenis and Rédei 2011, 2013). Also note that it is not known if the quantum analogues of (2) and (3) of Proposition 1 hold, i.e. it is not known whether every general probability space is extendable into a purely non-atomic one; it is conjectured that this is true.

4 Quantum Correlations and the Common Cause Principle

It is known that (relativistic) quantum field theory (QFT) predicts correlations between observables associated with spacelike separated hence causally independent spacetime regions such as spacelike separated double cones D_1 and D_2: this

is a consequence of violation of Bell's inequality in quantum field theory (Rédei and Summers (2002) reviews the most relevant facts about spacelike correlations in QFT). Causal completeness of QFT would therefore mean that the spacelike correlations have a common cause which is local in the sense that it is an observable (a projection) that belongs to an algebra of observables that is localized in the *intersection* of the backward light cones of the regions D_1 and D_2. "Common cause" in the previous sentence refers to the concept of common cause in the sense of Definition 2.

If the correlation is predicted by a faithful state, then there exist common cause projections in the local von Neumann algebra associated with a double cone region $D \supset D_1 \cup D_2$: this is entailed by Proposition 3 and the highly non-trivial feature of quantum field theory that the double cone algebras are type **III** von Neumann algebras and that the projection lattice of type **III** von Neumann algebras are atomless – so faithful states on type **III** von Neumann algebras define a measure theoretically purely nonatomic quantum probability space $(\mathcal{P}(\mathcal{N}), \phi)$, which are common cause complete by Proposition 3. It is *not* known however whether there exists common causes of the spacelike correlations that are localized in the *intersection* of D_1 and D_2. This problem has remained open since it was first formulated (Rédei 1997) (also see Chap. 12 in Rédei 1998). We only know that if a quantum field theory satisfies the Local Primitive Causality condition, then common causes exist that are localized in the *union* of the backward light cones of D_1 and D_2 (Rédei and Summers 2002, 2007, see also Chap. 8 in Hofer-Szabó et al. 2013). QFT is thus *weakly* causally complete and definitely *cannot* be considered as disconfirming evidence for the Common Cause Principle at this time.

It should be noted that *lattice* quantum field theory behaves differently from the perspective of causal completeness: since in lattice quantum field theory the local algebras of observables are finite dimensional matrix algebras and the quantum probability spaces that they determine are thus discrete (not purely nonatomic in the measure theoretic sense), lattice quantum field theory in $(1 + 1)$ dimension contains spacelike correlations that do not have common causes *at all* – local or no local (Hofer-Szabó and Vecsernyés 2012). This result motivates weakening of definition notion of common cause (Definition 2) by allowing the common cause to *not* commute with the commuting correlated projections. It seems that this weakening compensates for the measure theoretic discreteness of the non-classical probability spaces in lattice quantum field theory: lattice quantum field theory in $1 + 1$ dimension can be shown to be weakly causally complete with respect to non-commuting common causes (see Chap. 8 in Hofer-Szabó et al. 2013 for the details).

The notorious EPR correlations are typically considered as disconfirming evidence for the Common Cause Principle. The standard argument is that assumption of common causes of EPR correlations entails Bell's inequality, hence in view of violation of Bell's inequality one concludes that common causes of those correlations cannot exist. A careful look at the problem reveals however that problem is very subtle and that the standard arguments are too quick.

The first relevant observation is that before even raising the issue of possibility of common causes of EPR correlations, one has to set up a probabilistic model in which the EPR correlations are described – probability statements, such as events

being correlated or independent, are meaningful only within a fixed probabilistic model, not in general. It is non-trivial but provable that the EPR correlations can be represented by a classical probability measure space if the correlations are properly interpreted, i.e. if the values of quantum probabilities are considered as classical conditional probabilities, where the conditioning events are the measurement setups (see Szabó 2001; Rédei 2010; Hofer-Szabó et al. 2013). Once the probability models describing the EPR correlations have been set up, one can invoke the common cause extendability results presented in the previous section and conclude that the EPR correlations *can* in principle have common causes.

The second observation is therefore that one has to impose further conditions on the common cause models of the EPR correlations, conditions that express, in probabilistic terms, the causal relations that hold among the random events (including the random events representing the hypothetical common causes) by virtue of the fact that they have particular spatio-temporal locations. These are the "locality conditions". There are two types of locality conditions: those that are empirically testable because they involve observable random events (such as the outcome events and the events setting up the measuring device in a correlation experiment) – these are called "surface locality" conditions – and those that are *not* observed in the correlation experiment (such as the hypothetical common cause events and their occurrence in combination with observed events) – these are called "hidden locality conditions".

The third observation is that it turns out that in order to be able to derive a Bell-type inequality from the assumption of local common causes of EPR correlations, it is not sufficient to impose locality conditions: one also has to require "no conspiracy conditions". These latter requirements are statistical independence conditions expressing independence of the hypothetical common causes from the events that set up the measuring device in a fixed direction in the EPR correlation experiments. To complicate matters further, the "no conspiracy" conditions come in two forms: weak and strong. The difference is related to the fact that in an EPR correlation experiment one has more than one pair of correlations to explain in terms of common causes and each of the correlations can in principle have its own, distinct common cause that need not have anything to do with the common causes of other correlations. The *strong* no conspiracy condition demands then that *any combination* of the hypothetical common causes of the different pairs of correlated events occurring in an EPR correlation experiment is probabilistically independent of *any combination* of measurement setups measuring the correlations in question. Under these conditions one can prove the following.

Proposition 4 (Proposition 9.16, Hofer-Szabó et al. 2013). *There exist no hidden local, strongly non conspiratorial common cause models of all the EPR correlations: there exist four directions in the left, and four directions in the right wings of the EPR measurement setup such that the 16 correlations arising from measuring spin components in the 4×4 possible pairs of direction in the left and right wings, respectively do not all have a hidden local and strongly non conspiratorial common cause model.*

The proof of the above proposition is based on deriving Bell-type inequalities from the assumptions and showing that the inequality is violated by the 16 correlations (Hofer-Szabó et al. 2013, p. 170).

It must be emphasized that the assumptions in Proposition 4 *cannot* be weakened in the following sense: one can show that there *do* exist local (as opposed to *hidden* local), non-conspiratorial (as opposed to *strongly* non-conspiratorial) common cause explanations of EPR correlations (see Hofer-Szabó et al. (2013) and the references therein). The significance of this is that, since the hidden locality conditions involve non-observed events, the hidden locality condition is non-empirical, it is metaphysical in nature. Since it is needed to derive conditions (Bell's inequality) that are the basis on which one claims that the EPR correlations are disconfirming evidence for the metaphysical Common Cause Principle, this is in harmony with the spirit of the opening remark in this paper: metaphysical principles cannot be conclusively (dis)proved empirically – when assessing metaphysical principles, a large amount of epistemological modesty is required.

Acknowledgements Supported in part by the Hungarian Scientific Research Found (OTKA). Contract number: K100715.

References

Butterfield, J. 2007. Stochastic Einstein locality revisited. *The British Journal for the Philosophy of Science* 58: 805–867.

Cartwright, N. 2007. *Hunting causes and using them*. Cambridge: Cambridge University Press.

Gyenis, B., and M. Rédei. 2004. When can statistical theories be causally closed? *Foundations of Physics* 34: 1285–1303.

Gyenis, B., and M. Rédei. 2011a. Causal completeness of general probability theories. In *Probabilities, causes and propensities in physics*, Synthese library, vol. 347, ed. M. Suarez, 157–171. Dordrecht/New York: Springer.

Gyenis, B., and M. Rédei. 2011b. Causal completeness of probability theories – results and open problems. In *Causality in the sciences*, ed. P.M. Illari, F. Russo, and J. Williamson, 526–539. Oxford: Oxford University Press.

Gyenis, Z., and M. Rédei. 2011. Characterizing common cause closed probability spaces. *Philosophy of Science* 78: 393–409.

Gyenis, Z., and M. Rédei. 2013, forthcoming. Atomicity and causal completeness. *Erkenntnis*. doi:10.1007/s10670-013-9456-1.

Henson, J. 2005. Comparing causaility principles. *Studies in the History and Philosophy of Modern Physics* 36(3): 519–543.

Hofer-Szabó, G. 2008. Separate- versus *common*-common-cause-type derivations of the Bell inequalities. *Synthese* 163: 199–215.

Hofer-Szabó, G. 2011. Bell(δ) inequalities derived from separate common causal explanation of almost perfect EPR anticorrelations. *Foundations of Physics* 41: 1398–1413.

Hofer-Szabó, G., and P. Vecsernyés. 2012. Reichenbach's common cause principle in algebraic quantum field theory with locally finite degrees of freedom. *Foundations of Physics* 42: 241–255.

Hofer-Szabó, G., M. Rédei, and L. Szabó. 1999. On Reichenbach's common cause principle and Reichenbach's notion of common cause. *The British Journal for the Philosophy of Science* 50: 377–398.

Hofer-Szabó, G., M. Rédei, and L. Szabó. 2002. Common-causes are not common common-causes. *Philosophy of Science* 69(20): 623–636.

Hofer-Szabó, G., M. Rédei, and L. Szabó. 2013, forthcoming. *The principle of the common cause*. Cambridge: Cambridge University Press.

Hoover, K.D. 2003. Non-stationary time series, cointegration and the principle of the common cause. *The British Journal for the Philosophy of Science* 54: 527–551.

Kitajima, Y. 2008. Reichenbach's common cause in an atomless and complete orthomodular lattice. *International Journal of Theoretical Physics* 47: 511–519.

Marczyk, M., and L. Wrónski. 2013, forthcoming. A completion of the causal completability problem. *The British Journal for the Philosophy of Science*.

Mazzola, C. 2012a. Reichenbachian common cause systems revisited. *Foundations of Physics* 42: 512–523.

Mazzola, C. 2012b, forthcoming. Correlations, deviations and expectations: The extended principle of the common cause. *Synthese*. doi:10.1007/s11229-012-0089-8.

Portmann, S., and A. Wüthrich. 2007. Minimal assumption derivation of a weak Clauser–Horne inequality. *Studies in History and Philosophy of Modern Physics* 38: 844–862. Electronic preprint: http://www.arxiv.org/quant-ph/0604216.

Rédei, M. 1997. Reichenbach's common cause principle and quantum field theory. *Foundations of Physics* 27: 1309–1321.

Rédei, M. 1998. *Quantum logic in algebraic approach*, Fundamental theories of physics, vol. 91. Dordrecht/Boston: Kluwer

Rédei, M. 2010. Kolmogorovian censorship hypothesis for general quantum probability theories. *Manuscrito: Revista Internacional de Filosofia* 33: 365–380.

Rédei, M., and I. San Pedro. 2012. Distinguishing causality principles. *Studies in the History and Philosophy of Modern Physics* 43: 84–89. Preprint: http://philsci-archive.pitt.edu/9095/.

Rédei, M., and S. Summers. 2002. Local primitive causality and the common cause principle in quantum field theory. *Foundations of Physics* 32: 335–355.

Rédei, M., and S. Summers. 2007. Remarks on causality in relativistic quantum field theory. *International Journal of Theoretical Physics* 46: 2053–2062.

Reichenbach, H. 1956. *The direction of time*. Los Angeles: University of California Press.

Salmon, W. 1984. *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

San Pedro, I. 2008. The common cause principle and quantum correlations. Ph.D. thesis, Department of Philosophy, Complutensee University, Madrid.

Sober, E. 2008. *Evidence and evolution*. Cambridge: Cambridge University Press.

Sober, E., and M. Steel. 2012, forthcoming. Screening-off and causal incompleteness – a No-Go theorem. *The British Journal for the Philosophy of Science*.

Szabó, L. 2001. Critical reflections on quantum probability theory. In *John von Neumann and the foundations of quantum physics, institute vienna circle yearbook*, ed. M. Rédei and M. Stöltzner, 201–219. Dordrecht: Kluwer.

Wrónski, L. 2010. The common cause principle. Explanation via screeing off. Ph.D. thesis, Institute of Philosophy, Jagiellonian University, Cracow. Forthcoming as a book from Versita Publishing.

Wrónski, L., and M. Marczyk. 2010. Only countable Reichenbachian common cause systems exist. *Foundations of Physics* 40: 1155–1160.

Wrónski, L., and M. Marczyk. 2013, forthcoming. A new notioin of causal closedness. *Erkenntnis*. doi:10.1007/s10670-013-9457-0.

Wüthrich, A. 2004. *Quantum correlations and common causes. Bern studies in the history and philosophy of science*. Bern: Universität Bern.

A Note on Strong Causal Closedness and Completability of Classical Probability Spaces

Michał Marczyk and Leszek Wroński

1 Introduction

In recent years the status of Reichenbach's Common Cause Principle (Reichenbach 1971) and its various mathematical formulations have again become the subject of livid discussion. The principle, according to which for any correlated events there exists a common cause satisfying certain probabilistic conditions we discuss below (see Definition 4), has been extensively criticized both on the general philosophical level and more formally, especially with reference to quantum contexts. This did not stop it, however, from becoming one of the main inspirations for the causal Markov condition, making it basically a working assumption for the field of causal Bayesian nets. In the last decade a new direction of support for the principle appeared, in the way of formal results concerning causal closedness and completability of probability spaces.[1] It was proven e.g. that whenever a probability space contains a correlation for which no event fulfilling Reichenbach's conditions exists, the space in question can be extended to a bigger space (with the measure of the "old" events preserved) in which such an event does exist (Hofer-Szabó et al. 1999). An even stronger fact holds: the initial probability space can be extended to one containing common causes in the Reichenbachian sense for *all* correlations (for different variants of this claim and various arguments, see e.g. Marczyk and Wroński forthcoming; Gyenis and Rédei 2011; Wroński 2010).

[1] For a collection of references on both philosophical and formal issues related to the Common Cause Principle, see Wroński (2010).

M. Marczyk • L. Wroński (✉)
Department of Philosophy, Jagiellonian University, Grodzka 52, 31-044 Krakow, Poland
e-mail: michal.marczyk@gmail.com; leszek.wronski@uj.edu.pl

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_30,

In this technical note we show a still stronger fact. Common causes in the sense of Definition 4 can be classified into various types depending on the conditional probability they give to the correlated events. Various correlations admit various types (more on that below). It can be asked whether, if a correlation in a given space lacks a common cause of a certain type, the space can be extended (preserving measure) to a space in which the correlation possesses such a common cause. This has been answered in the affirmative in Hofer-Szabó et al. (1999). But we could go further: are there probability spaces such that *all* correlations in them contain common causes of all admissible types? (This is the problem of "strong causal closedness".) If so, is it always possible to extend the original space to such a space? (This is the problem of "strong causal completability".) Both of these problems were posed by M. Rédei during the "New Directions in the Philosophy of Science" conference in Bertinoro, October 17th–20th 2012; solving them also answers the first problem in Sect. 4 of Gyenis and Rédei (2013). It turns out the answer is "yes" on both counts.

In the sequel we first give the required definitions and formal problems, to then show the positive answers to the problems. We also ask a more general problem regarding the so called "common cause systems" and also answer it in the positive. Lastly, we provide a short discussion of a possible future area of research regarding the notion of a type of a common-cause-like explanation. For an extended philosophical discussion of related issues, see Wroński (2010) or Hofer-Szabó et al. (2013).

2 Basic Definitions and Results in the Literature

We shall start with a few basic concepts. In this paper we are dealing with classical probability spaces only. Since sample spaces are irrelevant for the topic of this note, our probability spaces will just be pairs of a Boolean σ-algebra $\mathcal{F}$ and a classical normalized measure P on $\mathcal{F}$. We will say that events $A, B \in \mathcal{F}$ are *(positively) correlated* if $P(A \cap B) > P(A)P(B)$.

The correlations which demand explanation by means of a common cause are of course those between factors which are not directly causally related to each other. We also feel no need to explain correlations between events which are *logically dependent*, that is, such that the occurrence (or nonoccurrence) of one logically implies the occurrence (or nonoccurrence) of the other.

We need thus to exclude from the class of all pairs of correlated events those in which the events stand in some relation. Since in general we do not know the causal situation we are dealing with, a minimal candidate for such a relation is that of logical independence. We could also go a bit further and consider only *genuinely* independent events; that is, events differing by more than a measure zero event. This is the topic of the following definition.

Definition 1 (Logical independence). We say that events $A, B \in \mathcal{F}$ are *logically independent* ($\langle A, B\rangle \in L_{ind}$) iff all of the following sets are nonempty:

- $A \cap B$;
- $A \cap B^{\perp}$;
- $A^{\perp} \cap B$;
- $A^{\perp} \cap B^{\perp}$.

We say that events $A, B \in \mathcal{F}$ are *logically independent modulo measure zero event*, or *genuinely independent* ($\langle A, B\rangle \in L^{+}_{ind}$), iff all of the following numbers are positive:

- $P(A \cap B)$;
- $P(A \cap B^{\perp})$;
- $P(A^{\perp} \cap B)$;
- $P(A^{\perp} \cap B^{\perp})$.

Equivalently, two events are genuinely independent if every Boolean combination of them has a non-zero probability of occurring. It is always true that $L^{+}_{ind} \subseteq L_{ind}$.

We will now provide the definitions needed to formulate the main concept, that is, the concept of a common cause for two correlated events, of which Reichenbach gave a precise probabilistic description.

Definition 2 (Screening off). Let $A, B \in \mathcal{F}$. An event C is said to be a *screener-off* for the pair $\{A, B\}$ if and only if $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$. In the case where A and B are correlated we also say that C *screens off* the correlation.

Definition 3 (Statistical relevance). Let $A, B \in \mathcal{F}$. We say that a family of events $\{C_i\}_{i\in I}$ satisfies the *statistical relevance* condition with regard to the pair $\{A, B\}$ if and only if whenever $i \neq j$ (for $i, j \in I$)

$$(P(A \mid C_i) - P(A \mid C_j))(P(B \mid C_i) - P(B \mid C_j)) > 0$$

Definition 4 (Statistical Common Cause). Let $\langle \mathcal{F}, P\rangle$ be a probability space. Let $A, B \in \mathcal{F}$. Any event $C \in \mathcal{F}$ different from both A and B[2] such that

- Both C and $C^{\perp}$ are screener-offs for the pair $\{A, B\}$,
- The pair $\{C, C^{\perp}\}$ satisfies the statistical relevance condition with regard to the pair $\{A, B\}$,

is called a *statistical common cause* (SCC) of A and B.

In this note we abstain from the popular practice of calling the above concept a "Reichenbachian Common Cause". For the reasons for this, see Wroński and Marczyk (2013).

[2]I.e. "Any event C which is not identical as a set to event A or event B".

Statistical common causes possess some features due to which they might seem to be reasonable candidates for explanations of correlations. First, notice that due to screening off, conditioning on a statistical common cause renders its effects statistically independent – the correlation disappears. Second, an SCC raises the probability of occurrence of both its effects, which is an initially natural, though controversial (see Beebee 1998), condition for causes in the context of probabilistic causation. Third, statistical common causes (as proved already by Reichenbach) have a certain deductive feature: namely, from both screening off conditions together with the statistical relevance condition it is possible to deduce the correlation between the effects of the cause.

The generalization of this concept from a simple "yes"/"no" event to something more akin to a random variable, corresponding to non-binary traits, is the notion of a "statistical common cause system", introduced (under a slightly different name) in Hofer-Szabó and Rédei (2004):

Definition 5 (Statistical Common Cause System). Let $\langle \mathcal{F}, P \rangle$ be a probability space. A partition of unity of $\mathcal{F}$ is said to be a *statistical common cause system* (SCCS) for A and B if and only if it satisfies the statistical relevance condition w.r.t. A and B, all its members are different from both A and B, and all its members are screener-offs for the pair.

The cardinality of the partition is called the *size* of the statistical common cause system.

Notice that SCCSs may be finite or at most countably infinite (Wroński and Marczyk 2010).

When we say that a probability space *contains* an SCCS, we mean that the SCCS is a partition of unity of the underlying algebra of the space. We also sometimes use the expression "common cause (system)" instead of "statistical common cause (system)" to shorten the formulation of some arguments.

SCCSs have similar explanatory virtues as SCCs: this is more or less evident with regard to the screening off and statistical relevance conditions, and it was shown in Hofer-Szabó and Rédei (2004) that existence of an SCCS for events $A, B \in \mathcal{F}$ entails a correlation between those events.

We now introduce the first of the notions connected with causal closedness of probability spaces, that is, with the existence of SCCs and SCCSs for all correlated pairs of events belonging to the given independence relation.

Definition 6 (Causal (up-to-)n-closedness). We say that a classical probability space is *causally (up-to-)n-closed*[3] w.r.t. to a relation of independence R_{ind} if and

[3]Notice that our n-closedness is a different notion from that of "causal N-completeness" employed by Gyenis and Rédei (2010). There the framework is that of general probability spaces and the correlations under investigation hold between random variables. Also, the choice of a specific notion of correlation is left open; whereas here, since we are talking about events, no such choice arises.

only if all pairs of correlated events independent in the sense of R_{ind} possess a proper statistical common cause or a proper statistical common cause system of size (at most) n.

If the space is causally 2-closed, we also say that it is *causally closed* or *common cause closed*.

If the space in question is not causally closed, it is natural to wonder whether it has a causally closed extension, namely a space which contains all the events of the original space (with the original measure) but is more "fine-grained", containing also the previously "missing" SCCs for the correlations.

Definition 7 (Extension). A classical probability space $\langle \mathcal{F}', P' \rangle$ is called an *extension* of the probability space $\langle \mathcal{F}, P \rangle$ iff there exists a Boolean algebra embedding h of $\mathcal{F}$ into $\mathcal{F}'$ such that for any $E \in \mathcal{F}$, $P(E) = P'(h(E))$.

Definition 8 (Causal ((up-to)-n)-completability). Suppose $\langle \mathcal{F}, P \rangle$ is a probability space and $\mathbf{G}$ is a family of pairs of correlated events which do not have an SCC in $\langle \mathcal{F}, P \rangle$. The space $\langle \mathcal{F}, P \rangle$ is *causally completable with regard to the family* $\mathbf{G}$ if there exists an extension $\langle \mathcal{F}', P' \rangle$ of $\langle \mathcal{F}, P \rangle$ by means of a homomorphism h which contains an SCC for $\langle h(A), h(B) \rangle$ for every pair $\langle A, B \rangle \in \mathbf{G}$.

The space $\langle \mathcal{F}, P \rangle$ is *causally n-completable with regard to the family* $\mathbf{G}$ if there exists an extension $\langle \mathcal{F}', P' \rangle$ of $\langle \mathcal{F}, P \rangle$ by means of a homomorphism h which contains an SCCS of size n for $\langle h(A), h(B) \rangle$ for every pair $\langle A, B \rangle \in \mathbf{G}$.

The space $\langle \mathcal{F}, P \rangle$ is *causally up-to-n-completable with regard to the family* $\mathbf{G}$ if there exists an extension $\langle \mathcal{F}', P' \rangle$ of $\langle \mathcal{F}, P \rangle$ by means of a homomorphism h which contains an SCCS of size up to n for $\langle h(A), h(B) \rangle$ for every pair $\langle A, B \rangle \in \mathbf{G}$.

The next definition to be introduced is that of an atomless probability space. It turns out that such spaces are extremely rich in statistical common causes (see Fact 2) and allow the proof of many results concerning causal completability.

Definition 9 (Atomless probability space). A probability space $\langle \mathcal{F}, P \rangle$ is *atomless* if for any $C \in \mathcal{F}$, if $P(C) > 0$, then there exists $D \in \mathcal{F}$ such that $D \subset C$ and $0 < P(D) < P(C)$.

Remarkably, this definition suffices to prove the following fact, sometimes referred to as "denseness property":

Fact 1. *If C is an event of positive probability r in an atomless probability space $\langle \Omega, \mathcal{F}, P \rangle$, then for any real number s such that $0 < s < r$ there exists an event $D \in \mathcal{F}$ such that $D \subset C$ and $P(D) = s$.*

See e.g. p. 46 of Fremlin (2001) for a proof.

Concerning the aforementioned richness of atomless spaces, the following

Fact 2 (Gyenis and Rédei 2004). *All atomless probability spaces are causally closed w.r.t. L_{ind}.*

together with the fact that all classical probability spaces can be extended to atomless spaces (Wroński 2010; Marczyk and Wroński forthcoming) leads to the conclusion that *all classical probability spaces can be extended to causally closed spaces*.

Moreover, Marczyk and Wroński (forthcoming) prove a variety of results on atomless spaces: for example, for any natural number n, for each correlation between logically independent events these spaces contain uncountably many SCCSs of size n for that correlation. And so they are causally n-closed for any n, which means all classical probability spaces are (for any n) causally n-completable (and so, trivially, up-to-n-completable) w.r.t. L_{ind}.

3 New Results on Strong Causal Closedness and Completability

The Reader may be already convinced that for those who seek a Reichenbach-style explanation of correlations, atomless probability spaces form a veritable paradise, whether one looks for SCCs or SCCSs of any particular size. However, it is possible to show even more. Suppose in some probability space with the measure P events A and B are correlated and C is their statistical common cause. The quintuple $\langle P(C), P(A|C), P(A|C^{\perp}), P(B|C), P(B|C^{\perp})\rangle$ is called the *type* of C; this notion was introduced in Hofer-Szabó et al. (1999), where the Authors also derive, assuming that A and B are correlated events, the conditions of admissibility for a quintuple of real numbers to serve as appropriate conditional probabilities for some SCC of A and B; whether such an SCC exists in the given space is another matter, of course. The main idea is that a single correlation can possess SCCs of various types, which invites a new notion of causal closedness and completability. The first part of the following definition comes from Hofer-Szabó et al. (1999):

Definition 10 (Types of SCCs and SCCSs). In a probability space with measure P, a statistical common cause C of a correlation between events A and B is said to have (be of) the type $\langle r_C, r_{A|C}, r_{B|C}, r_{A|C^{\perp}}, r_{B|C^{\perp}}\rangle$ if each of these numbers is equal to the probability indicated by its index; e.g. if $r_{A|C} = P(A|C)$ and so on.

In a probability space with measure P, a statistical common cause system $\{C_i\}_{i\in I}$ of a correlation between events A and B is said to have (be of) the type $\langle\{r_{C_i}\}_{i\in I}, \{r_{A|C_i}\}_{i\in I}, \{r_{B|C_i}\}_{i\in I}\rangle$ if each of these numbers is equal to the probability indicated by its index; e.g. if for any $j \in I$ $r_{A|C_j} = P(A|C_j)$ and so on.

Notice that in the part of the definition concerning SCCSs we need to include *all* numbers $\{r_{C_i}\}_{i\in I}$, while in the part concerning SCCs we only included r_C (and not $r_{C^{\perp}}$). This is because $r_{C^{\perp}} = 1 - r_C$, so the information about $r_{C^{\perp}}$ is superfluous. There is no convenient analogue of this feature when we move to possibly infinite SCCSs.

As mentioned above, precise conditions for SCC admissibility for a given correlation were derived in Hofer-Szabó et al. (1999). These follow directly from Definition 4. It is immediate that the notion of admissibility of types makes sense also in the case of SCCSs of any cardinality.

Definition 11 (Strong causal closedness/completability). A probability space is strongly causally closed w.r.t. a relation of logical independence R_{ind} if for any correlated pair in R_{ind}, and for any type T admissible for that pair, it contains an SCC of type T.

A probability space is strongly causally completable w.r.t. R_{ind} if it can be extended to a space which is strongly causally closed w.r.t. R_{ind}.

Z. Gyenis and M. Rédei asked the following problem[4]: are any spaces strongly causally closed? If so, are any spaces strongly causally completable? The answer is "yes" on both counts. We should not be surprised that again it is the atomless spaces which are doing the job:

Theorem 1. *All atomless classical probability spaces are strongly causally closed w.r.t.* L^{+}_{ind}. *All classical probability spaces are also strongly causally completable w.r.t. the same relation.*

Proof. Consider an atomless probability space with measure P, two genuinely independent correlated events A and B in it, and an SCC type T admissible for that correlation. The type T is given by the probabilities $P(C)$, $P(A|C)$, $P(A|C^{\perp})$, $P(B|C)$, $P(B|C^{\perp})$ for a statistical common cause C of that type. This in turn determines the probabilities $P(C \cap A \cap B)$, $P(C \cap A^{\perp} \cap B)$, $P(C \cap A \cap B^{\perp})$ and $P(C \cap A^{\perp} \cap B^{\perp})$ a common cause C of type T for A and B would have.

We can therefore select four events with the appropriate probabilities ($P(C \cap A \cap B)$ etc.) as subsets of $A \cap B$, $A \cap B^{\perp}$, $A^{\perp} \cap B$ and $A^{\perp} \cap B^{\perp}$ accordingly. These events exist because the space, due to being atomless, has the denseness property. Now we construct the event C as the union of these four events. It is immediate that C is a statistical common cause of type T for the original correlated pair. Since the choice of a correlated pair and admissible type was fully general, we have proved that all atomless probability spaces are strongly causally closed w.r.t. L^{+}_{ind}.

For the last step, since all classical probability spaces can be extended to atomless classical probability spaces, all classical probability spaces are also strongly causally completable. **Q.E.D.**

Notice that nothing essential for the proof depended on that we have been concerned with SCCs and not (at least finite) SCCSs. Indeed, the proof carries over to the more general case, which requires new definitions which are to be seen as natural generalisations of those gathered in Definition 11.

Definition 12 (Strong causal n-closedness/n-completability). A probability space is strongly causally n-closed w.r.t. a relation of logical independence R_{ind} if for any correlated pair in R_{ind} and for any type T for an SCCS of size n which is admissible for that pair, it contains an SCCS of size n of type T.

[4]M. Rédei gave the definitions of strong causal closedness and completability as well as stated the problems during the "New Directions in the Philosophy of Science" conference in Bertinoro, October 17th–20th 2012; a similar problem is the first one in Sect. 4 of Gyenis and Rédei (2013).

A probability space is strongly causally n-completable w.r.t. R_{ind} if it can be extended to a space which is strongly causally n-closed w.r.t. R_{ind}.

We omit the definitions for strong up-to-n-closedness and completability since they are obvious and the paper already contains too much notation.

Now, as mentioned earlier, nothing essential in the proof of Theorem 1 depends on us being concerned with SCCs and not finite SCCSs. It is just a matter of finding more subsets of $A \cap B$, $A \cap B^{\perp}$, $A^{\perp} \cap B$, and $A^{\perp} \cap B^{\perp}$, with the probabilities easily calculable from the type. Therefore we omit a separate proof of the following:

Theorem 2. *All atomless classical probability spaces are strongly causally n-closed w.r.t. L^{+}_{ind}. All classical probability spaces are also strongly causally n-completable w.r.t. the same relation.*

We will say a probability space is strongly causally ω-closed w.r.t. R_{ind} iff all pairs belonging to R_{ind} possess countably infinite SCCSs of all admissible types. Define a space as strongly causally ω-completable w.r.t. R_{ind} iff it can be extended to a space which is strongly causally ω-closed w.r.t. R_{ind}. Then the following proposition also follows; here sums of probabilities become series of positive terms and it is still possible to carry out essentially the same construction as in the proof of Theorem 1. A detailed proof will be presented elsewhere.

Proposition 1. *All atomless classical probability spaces are strongly causally ω-closed w.r.t. L^{+}_{ind}. All classical probability spaces are also strongly causally ω-completable w.r.t. the same relation.*

References

Beebee, H. 1998. Do causes raise the chances of effects? *Analysis* 58(3): 182–190.

Fremlin, D. 2001. *Measure theory*, vol. 2. Colchester: Torres Fremlin.

Gyenis, B., and M. Rédei. 2004. When can statistical theories be causally closed? *Foundations of Physics* 34(9): 1284–1303.

Gyenis, B., and M. Rédei. 2010. Causal completeness in general probability theories. In *Probabilities, causes, and propensities in physics*, Synthese library, vol. 347, ed. M. Suárez, 157–171. Dordrecht: Springer

Gyenis, Z., and M. Rédei. 2011. Characterizing common cause closed probability spaces. *Philosophy of Science* 78: 393–409.

Gyenis, Z., and M. Rédei. 2013. Atomicity and causal completeness. *Erkenntnis*. doi:10.1007/s10670-013-9456-1.

Hofer-Szabó, G., and M. Rédei. 2004. Reichenbachian common cause systems. *International Journal of Theoretical Physics* 43(7/8): 1819–1826.

Hofer-Szabó, G., M. Rédei, and L.E. Szabó. 1999. On Reichenbach's common cause principle and Reichenbach's notion of common cause. *The British Journal for the Philosophy of Science* 50(3): 377–399.

Hofer-Szabó, G., M. Rédei, and L.E. Szabó. 2013. *The common cause principle*. Cambridge: Cambridge University Press.

Marczyk, M., and L. Wroński. forthcoming. A completion of the causal completability problem. *The British Journal for the Philosophy of Science*.

Reichenbach, H. 1971. *The direction of time*. Berkeley/Los Angeles/London: University of California Press. Reprint of the 1956 edition.

Wroński, L. 2010. The common cause principle. Explanation via screening off. Ph.D. thesis, Jagiellonian University, Kraków, archived at jagiellonian.academia.edu/LeszekWroński. Forthcoming as a book from Versita Publishing.

Wroński, L., and M. Marczyk. 2010. Only countable Reichenbachian common cause systems exist. *Foundations of Physics* 40: 1155–1160.

Wroński, L., and M. Marczyk. 2013. A new notion of causal closedness. *Erkenntnis*. doi:10.1007/s10670-013-9457-0.

Artificial Examples of Empirical Equivalence

Pablo Acuña

1 Three Sources of Empirical Equivalence

The problem of empirical equivalence (EE) and underdetermination (UD) of theory choice can be expressed by means of a simple argument. The first premise states that *for any theory T that entails the class of observational consequences O there is another theory T′ whose class of observational consequences is also O*. The second premise is that *entailment of evidence is the only epistemically justified criterion for the confirmation of theories*. From these two premises it follows that the objectivity – and maybe even the rationality – of theory choice is threatened. Notice that the universal scope of the first premise implies that the problem holds for science as a whole, in the sense that *all* theories are affected by EE and UD.

EE between theories can be instantiated in four different ways: *(i)* by algorithms, *(ii)* by accommodating auxiliary hypotheses according to the Duhem-Quine thesis, *(iii)* by the regular practice of science, and *(iv)* by concrete artificial examples. The universal scope of the first premise of the problem is supported by *(i)* and *(ii)*. If there exist algorithms that are able to produce EE theories given any theory *T*, or if it is always possible to accommodate evidence by means of manipulation of auxiliary hypotheses, then it follows that EE is a condition that holds for any theory whatsoever. Elsewhere I have argued that neither *(i)* nor *(ii)* really work as possible sources of EE.[1] In the case of *(iii)*, Larry Laudan and Jarret Leplin proposed a twofold way out of the problem. First, they claim that EE is

[1](Acuña and Dieks 2013).

P. Acuña (✉)
Institute for History and Foundations of Science, Utrecht University, Budapestlaan 6, 3584 CD Utrecht, The Netherlands
e-mail: p.t.acunaluongo@uu.nl

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_31,

a time-indexed feature – in the sense that it is a condition essentially relative to a specific state of science and technology – and that it might get broken by future scientific or technological developments. Second, Laudan and Leplin argue that the UD between EE theories can be broken by means of non-consequential empirical evidence – even if the predictive equivalence remains.[2]

In this paper I will tackle the remaining source of EE, namely, concrete examples of artificially generated pairs of empirically equivalent theories. These examples are neither the outcome of the application of algorithms, nor obtained by manipulation of auxiliary hypotheses given an actual theory *T*. They are not the result of the practice of real science either. Rather, they have been *cooked up* and exploited by philosophers of science in order to speculate about their epistemological consequences. I will address an examination of three examples of artificially generated EE theories that have received attention in the philosophy of science literature: Bas van Fraaseen's alternative formulations of Newton's mechanics; the theories involved in the Poincaré-Reichenbach "parable"; and the case of predictively equivalent *total theories* or *systems of the world*.

2 Van Fraassen's Alternative Formulations of Newton's Theory

In *The Scientific Image* Bas van Fraassen introduced an argument for his constructive empiricism that involves an example of EE. He presents Newton's theory as a theory about the motion of bodies in space and the forces that determine such motions. The crucial feature that grounds van Fraassen's argument is that Newton's theory is supposed to be committed to the view that physical objects exist in absolute space. Thus, by reference to absolute space the concepts of absolute motion and absolute velocity become meaningful. Then, van Fraassen proposes

> let us call Newton's theory (mechanics and gravitation) *TN*, and $TN_{(v)}$ the theory *TN* plus the postulate that the center of gravity of the solar system has constant absolute velocity *v*. By Newton's own account, he claims empirical adequacy for $TN_{(0)}$; and also that if $TN_{(0)}$ is empirically adequate, then so are all the theories $TN_{(v)}$. (Van Fraassen 1980, p. 46)

Newton's most famous argument for the existence of absolute space is given by the thought experiment of the rotating bucket. In order to make sense of the acceleration of the rotating water in the bucket, the reality of absolute space has to be asserted, Newton argued. Van Fraassen's line of reasoning is that if absolute space exists, as Newton believed, then the concept of absolute motion of objects in space gets defined and so does the concept of absolute velocity. However, since – unlike absolute acceleration – absolute velocity has no observable effects, there

[2](Laudan and Leplin 1991).

are infinitely many predictively equivalent rival formulations of *TN*, each of them assigning a different specific value to the absolute velocity of the solar system's center of gravity.

According to van Fraassen, this entails a problem for the realist. The realist is committed to the view that only *one* of these alternative formulations is the true theory, but the realist's choice cannot be determined on evidential grounds.[3] For the constructive empiricist, van Fraassen argues, there is no such problem. In his/her case there is no commitment to the truth of the theory, but only to its empirical adequacy. Therefore, for the constructive empiricist it is enough to accept the empirical content of the theory as empirically adequate and assume a dodging attitude with respect to its non-empirical content – including the value for the absolute velocity of the solar system, of course. In other words, the empirical equivalence of the alternative formulations of Newton's theory does not necessarily put the constructive empiricist in the position of having to make a choice.[4]

A systematic consideration of van Fraassen's challenge shows that the real problem is not EE. It is true that Newton endorsed absolute space and that his preferred alternative was $TN_{(0)}$. However, rather than a case of EE, what is behind van Fraassen's example is a situation where there is a superfluous hypothesis within *TN*. A hypothesis is superfluous if it is not logically relevant for the derivation of any empirical consequences of the theory it forms a part of; and a hypothesis being superfluous is a strong indication that it represents nothing physical – an ontologically empty hypothesis, we could say. Therefore, the fact that the predictive equivalence between van Fraassen's alternative formulations is grounded on the stipulation of a specific value for a superfluous parameter – absolute velocity – indicates that we have a problem with the foundations of $TN_{(v)}$, rather than a genuine case of EE.

The superfluity problem with the concept of absolute velocity in Newton's theory has actually been solved and, *a fortiori*, the specious problem of EE gets dissolved. The key concept is a structure known as *neo-Newtonian spacetime*.[5] The basic elements of this structure are event-locations – the spatiotemporal locations where physical events (can) occur. A temporal separation – that can be zero – is defined

[3] From the viewpoint of the semantic conception of scientific theories, that van Fraassen endorses, the realist is committed to the view that *only one of the models that satisfy* $TN_{(v)}$ *correctly represents the world*. In the case of Newton, that model is given by $TN_{(0)}$, though the absolute velocity of the solar system is not a phenomenon.

[4] In semantic terms, the constructive empiricist stance is that to accept $TN_{(v)}$ as empirically adequate means that $TN_{(v)}$ *has a model which is empirically adequate*, i.e., it possesses an empirical substructure isomorphic to all phenomena. Making a choice is possible for a constructive empiricist, and he/she could do it based on pragmatic features of one of the alternative formulations. However, such a choice does not have an epistemic import, according to van Fraassen's view.

[5] Neo-Newtonian spacetime is the result of the work of P. Frank in 1909, and E. Cartan and K. Friedrichs in the 1920s. For a technical exposition of neo-Newtonian space-time and references to the seminal works of Frank, Cartan and Friedrichs, see (Havas 1964). For simpler expositions see (Sklar 1974), and (Stein 1970).

for all pairs of event-locations, and this is an absolute relation in the sense that it is not relative to particular frames of reference, states of motion, etc. A class of simultaneous event-locations – those for which their temporal separation is zero – forms a *space,*[6] and the structure of each space is that of Euclidean three-dimensional space.

The feature that differentiates Newtonian absolute space and neo-Newtonian spacetime is the way in which the spaces are connected or "glued-together". In absolute Newtonian space points conserve their spatial identity through time, and it is thus meaningful to ask whether a certain point or event-location at time t_1 is identical with some point or event-location at time t_2. In neo-Newtonian spacetime this question makes no sense, since the notion of spatial coincidence is only defined for simultaneous event-locations.

This difference in structure has a straightforward effect on the way that velocity is defined in each case. In neo-Newtonian spacetime it is coherent to ask for the velocity of a particle between two events in its history, but only if we are talking about its velocity with respect to some particular object – we can ask if the distance of the particle with respect to another object is the same as its distance to that same object at an earlier time, of course. But since absolute spatial coincidence through time is not defined, the concept of "absolute velocity" is meaningless in neo-Newtonian spacetime. Since points or event-locations do not conserve their identity through time, we cannot ask if the distance of an object with respect to a certain point in space at time t_2 has changed, or not, with respect to the distance between the object and that same point at an earlier time t_1.

Even though "absolute position" and "absolute velocity" are undefined, the concept of "absolute acceleration" is well defined in neo-Newtonian space-time, but this definition does not require reference to absolute space. First we need to introduce the three-place relation of "being inertial" between three non-simultaneous event-locations a, b and c. The relation holds if there is a possible path for a particle such that three events in its history are located at a, b and c, and if the particle is at rest in some inertial frame – a frame in which no inertial forces act upon any physical system at rest in it. More generally, a collection of events conforms an inertial class of events if they are all locations of events in the history of some particle that moves free of forces, a particle that moves *inertially*.

We can now explain the absolute acceleration of a particle along a time interval. Take the particle at the beginning of the interval and find an inertial frame in which the particle is at rest. At the end of the interval we find the new inertial frame in which the particle is at rest. Then we find the relative velocity of the second frame with respect to the first one at the end of the interval. Even though there

[6]In neo-Newtonian space-time simultaneity is an equivalence relation: every event is simultaneous with itself, if a is simultaneous with b then b is simultaneous with a, and if a is simultaneous with b and b is simultaneous with c then a is simultaneous with c. Therefore, it is possible to divide the class of all events in equivalence classes under the relation of simultaneity – classes that have no members in common and that taken together exhaust the class of all events.

is no such thing as the absolute velocity of the first inertial frame, we do know that, by definition, its velocity – with respect to any other inertial frame – has not *changed* throughout the interval. Therefore, the relative velocity of the second frame with respect to the first one gives us the absolute change of velocity throughout the interval, for the particle was at rest with respect to the first frame at the initial instant, and at rest with respect to the second frame at the end. We take this absolute change of velocity and divide it by the time separation between the initial and final event-locations and we obtain the absolute acceleration of the particle over the interval, and by applying the usual limiting process of differential calculus on the time interval we generate the concept of instant absolute acceleration. That is, absolute acceleration, within the context of a neo-Newtonian space-time, is defined not as relative to absolute space, but as relative to any inertial frame.[7]

Now we can go back to van Fraassen's challenge. As I mentioned above, the formulation of Newtonian mechanics in terms of neo-Newtonian spacetime can be understood as the solution for an unease with its foundations: the superfluous concept of absolute velocity. That is, the example that van Fraassen offers is not a genuine case of EE between rival theories. The problem is simply that the presence of the superfluous parameter v in *TN manifested* in that alternative, apparently incompatible formulations could be given. Neo-Newtonian spacetime solves this problem. It allows a more satisfactory formulation of *TN* in which the superfluous parameter has been swept away, so that there is no EE arising from different values assigned to v. In other words, the EE equivalence between van Fraassen's formulations was not the sickness, but just a symptom. Therefore, van Fraassen's challenge cannot be fruitfully used in order to extract conclusions related to the problem of EE and UD.[8] These remarks, of course, do not intend a refutation of constructive empiricism. The point is only that this particular example has no relevant consequences regarding the problem of EE and UD.

[7]Notice that the formulation of Newton's theory in terms of neo-Newtonian spacetime does not imply a rejection of a substantivalist position. Spacetime can still be postulated as the *arena* in which physical events occur, and the substantivalist can still argue that such arena possesses an independent existence, not reducible to relations between physical objects. See (Earman 1970).

[8]The reader might complain that since the alternative formulations of $TN_{(v)}$ are based on a theory that forms part of "real" physics means that van Fraassen's argument is a case in which EE is supposed to arise from the actual practice of science, not an artificial example. However, notice that a choice between formulations of $TN_{(v)}$ was never an issue for the scientific community, there never was a scientific debate about what is the correct value of v. What did happen was a debate concerning the meaningfulness of v – Leibniz's arguments in the *Leibniz-Clarke correspondence*, for example. This debate was not grounded on a problem of EE and UD of theory choice, it was (is) a debate about the ontology of space. This is yet another indication that van Fraassen is exploiting a problem with the foundations of Newton's theory in order to create a (specious) artificial case of EE.

3 The Poincaré-Reichenbach Argument

Henri Poincaré (1952) introduced a famous argument for the conventionality of geometry that has been considered as an example of EE. He designed a "parable" in which a universe given by a Euclidean two-dimensional disk is inhabited by flatlanders-physicists. The temperature on the disk is given by $T(R^2 - r^2)$, where R is the radius of the disk and r is the distance of the location considered to the center of the disk – therefore, the temperature at the center of the disk is TR^2 and at the edge it is 0 ° absolute. The inhabitants of this world are equipped with measuring rods that contract uniformly with diminishing temperatures, and all such rods have length 0 when their temperature is 0 °. The two-dimensional physicists proceed to measure distances in the disk with their rods in order to determine the geometry of their world; but they assume, falsely, that the length of their rods remains invariant upon transport – the flatlanders themselves also contract with diminishing temperature. Accordingly, the result they obtain is that they live in a Lobachevskian plane of infinite extent. For example, they measure that the ratio of a circumference to its radius is always greater than 2π. They obtain the same result by using measurements performed with light rays, for their universe is characterized by a refraction index $1/(R^2 - r^2)$; but they falsely assume that light beams travel along geodesics in their world, and that the index of refraction of vacuum is everywhere the same.

The parable also tells us that one particularly smart and revolutionary scientist in the disk comes up with the correct theory about the geometry of their world. Even though they are not able to observe effects of the temperature gradient $(R^2 - r^2)$ and of the refraction index $1/(R^2 - r^2)$, our brilliant physicist notices that, by assuming the reality of such unobservable features, the result is that the geometry of their universe is that of a finite Euclidean disk. The scientific community on the disk does not have the resources to make an evidentially based decision between the theories, and Poincaré's point is that the only way they can determine a specific geometry for their world is in terms of a *convention*. Poincaré also states that in our three-dimensional world we are, in principle, in the same situation. Empirically equivalent theories of our world that differ in the geometry they pose are analogously attainable. Therefore, the geometry of the physical world is a matter of convention also for us.

Two remarks can be made at this point about Poincaré's argument. First, it is clear that it is not an argument directly aiming to extract conclusions about the problem of EE and UD, it is an argument concerning the epistemology of geometry. This feature indicates that if we are going to take it as a concrete example of EE and UD some provisos must be introduced. Second, it is also clear that the example of empirically equivalent theories it considers is of a peculiar kind. The theories are not about the "real" physical world. The universe of the flat disk is a mental construction and, as such, it can be arranged and manipulated so that it totally complies with the description given by each of the theories. The world described by the theories is an ad hoc world. But this feature of the argument suggests that the example of EE involved is not a very serious or threatening one. The choice between the theories

is underdetermined because the whole situation can be conceptually manipulated in the required way.

Hans Reichenbach, in *The Philosophy of Space and Time*, introduced a sort of generalization of the argument. He presented it as a theorem showing that from any spacetime theory about the *real* physical world it is possible to obtain an alternative theory which is predictively equivalent but that assigns a different geometry:

> Mathematics proves that every geometry of the Riemannian kind can be mapped upon another of the same kind. In the language of physics this means the following:
>
> *Theorem* θ: 'Given a geometry G' to which the measuring instruments conform, we can imagine a universal force F which affects the instruments in such a way that the actual geometry is an arbitrary geometry G, while the observed deviation from G is due to a universal deformation of the measuring instruments'. (Reichenbach 1958, pp. 32–33)[9]

Under this formulation, the argument for the conventionality of geometry has a more substantial upshot on the problem of EE and UD. Reichenbach claims that the parable that Poincaré introduced can be effectively applied to "real" spacetime theories. For example, it could be stated that general relativity is empirically equivalent to a Newtonian-like theory of gravitation in which the curvature of spacetime is replaced by the action of a universal force. This complies with the first remark I made above regarding Poincaré's parable. Under Reichenbach's formulation, the argument for the conventionality of geometry can, in principle, be considered as an instance of EE involving theories about *our* world.

However, we still need to be precise about in what sense this argument, that primarily concerns the epistemology of geometry, affects the problem of EE and UD. For this purpose it is useful to take a look at what exactly Reichenbach is arguing for. The conventionalist stance he defends is weaker than Poincaré's. According to Reichenbach, what is a matter of convention regarding geometry are not, bottom line, the geometric features of the physical world, but the specific "language" in which those features are expressed. This argument relies on the concept of *coordinative definition*, that is, arbitrary definitions that settle units of measurement and which ground the particular conceptual systems that underlie physical theories:

> Physical knowledge is characterized by the fact that concepts are not only defined by other concepts, but are also coordinated to real objects. This coordination cannot be replaced by an explanation of meanings, it simply states that *this concept* is coordinated to *this particular thing*. In general this coordination is not arbitrary. Since the concepts are interconnected by testable relations, the coordination may be verified as true or false, if the requirement of uniqueness is added, i.e., the rule that the same concept must always denote the same object. The method of physics consists in establishing the uniqueness of this coordination, as Schlick has clearly shown. But certain preliminary coordinations must be determined before the method of coordination can be carried any further; these first coordinations are therefore definitions which we shall call *coordinative definitions*. They

[9] A *universal force*, roughly speaking, is a force that acts equally on all physical objects and that it cannot be shielded against. A *differential force*, on the contrary, can be shielded against and does not act equally on all physical objects. See (Reichenbach 1958, §6).

> are *arbitrary*, like all definitions; on their choice depends the conceptual system which develops with the progress of science.
>
> Wherever metrical relations are to be established, the use of coordinative definitions is conspicuous. If a distance is to be measured, the unit of length has to be determined beforehand by definition. This definition is a coordinative definition. (Reichenbach 1958, pp. 14–15)

Now it becomes clear why I said that Reichenbach's conventionalist view is a "weak" one. What is at stake in the EE between theory $T = F + G$ and $T' = F' + G'$ – where F denotes the set of forces that affect physical objects according to T, and T' is that same set plus a universal force T' that accounts for the deviation from geometry G according to T' – is only a divergence regarding the particular coordinative definitions that are presupposed by the theories. That is, we are in a situation analogous to a decision concerning whether Lionel Messi's height is 1.69 m or 5 ft and 7 in. In the case of Poincaré's disk, there are two different coordinative definitions at stake: one states that distances measured by rods have to be corrected according to a certain law, whereas in the other the measuring rods are rigid bodies that always express correct distances. Reichenbach's view on the conventionality of geometry is "linguistic", we could say. T and T' are two versions of the same theory expressed in different geometrical *languages*. To state that T is truer or more correct than T', or vice versa, is analogous to say that "meter" is a more correct unit of measurement than "foot".[10]

If Reichenbach is right, then the case of EE between T and T' that the argument involves is a harmless one. The choice between the theories is just a matter of the language we pick to express the same physical theory. Under Reichenbach's view

[10]Reichenbach also argues that the *default* language is the geometry in which universal forces are set to the zero value. If we do so, then the question regarding the specific geometry of the physical world becomes really meaningful, not only a matter of linguistic definitions: "The forces which we called universal are often characterized as forces *preserving coincidences*; all objects are assumed to be deformed in a way that the spatial relations of adjacent bodies remain unchanged. [. . .] It has been correctly said that such forces are not demonstrable, and it has been correctly inferred that they have to be set equal to zero if the question concerning the structure of space is to be meaningful. It follows from the foregoing considerations that this is a *necessary* but not a *sufficient* condition. Forces *destroying coincidences* must also be set equal to zero, if they satisfy the properties of the universal forces [. . .]; only then is the problem of geometry uniquely determined. [. . .]

We can define such forces as equal to zero because a force is no absolute datum. When does a force *exist*? By force we understand something which is responsible for a *geometrical change*. If a measuring rod is shorter at one point than at another, we interpret this contraction as the effect of a force. The existence of a force is therefore dependent on the coordinative definition of geometry. If we say: actually a geometry G applies, but we measure a geometry G', we define at the same time a force F which causes the difference between G and G'. The geometry G constitutes the zero point for the magnitude of a force. If we find that there result several geometries G' according as the material of the measuring instrument varies, F is a differential force; in this case we gauge the effect of F upon the different materials in such a way that all G' can be reduced to a common G. If we find, however, that there is only one G' for all materials, F is a universal force. In this case we can renounce the distinction between G and G', i.e., we can identify the zero point with G', thus setting F equal to zero. This is the result that our definition of the rigid body achieves" (Reichenbach 1958, pp. 27–28).

the conventionality of geometry has no special upshot on the problem of EE and UD as defined above. It is true that the choice between T and T' can be done only in terms of pragmatic considerations such as simplicity – empirical evidence, by definition, cannot settle the case. However, this is not a scientific or epistemological problem at all, for the choice does not involve incompatible rivals that differ in the way they describe the world. If we follow Reichenbach's line of thought, a genuine case of EE and UD would happen only if the theories involved postulate incompatible geometrical features for the world *provided that in both theories the universal forces are set to the zero value*. There is nothing in Reichenbach's argument to believe that this cannot happen, but it does not involve any example of this kind either.

This easy way out of the problem works only if Reichenbach is right, of course. His position regarding the epistemology of geometry is, clearly, quite close to the verificationist criterion of meaning endorsed by most of logical positivists. As it is known, this criterion has been shown to be untenable, and Reichenbach's view of the meaning of geometrical statements as reducible to coordinative definitions falls prey, *mutatis mutandis*, to the typical objections that have been leveled against logical positivistic semantics. That is, there are good reasons to think that Reichenbach's position is wrong, and, *a fortiori*, that the case of EE involved in his argument might be a relevant example with respect to the problem of UD of theory choice.

However, it turns out that even if we consider the case of $T = F + G$ vs. $T' = F' + G'$ as a genuine case of EE, this does not necessarily imply that we are dealing with a case of UD. The reason is given by the evidential status of the "universal forces". We can understand Reichenbach's theorem as stating that spacetime theories can have alternative empirically equivalent formulations by means of universal forces, and we can assume – unlike Reichenbach – that such alternatives are *genuine* rivals. However, that there exists an EE rival that postulates the reality of universal forces is not, *ipso facto*, an indication that the choice to be made is underdetermined by the empirical evidence. All "real" physical theories that invoke forces as the cause for dynamical effects postulate these forces as associated to *observable* effects; but the universal forces involved in Reichenbach's arguments are not at all like these "typical" forces. They are, in principle, not associated to any empirically detectible effect. The reality of usual, differential forces in physical theories is evidentially supported by the observable effects they cause, but this is not the case with universal ones. That is, in the case of $T' = F' + G'$ there is a hypothesis which is not, in principle, evidentially warranted. Therefore, we can conclude that $T = F + G$ possesses a higher degree of evidential support than T'. As Richard Boyd states it:

> Even though "$F \,\&\, G$" and "$F' \,\&\, G'$" have the same observational consequences (in the light of currently accepted theories), they are not equally supported or disconfirmed by any possible experimental evidence. Indeed, *nothing* could count as experimental evidence for "$F' \,\&\, G'$" in the light of current knowledge. This is so because the [universal] force f' required by F' [the class of the forces postulated by our T'] is dramatically unlike those

> forces about which we now know – for instance, it fails to arise as the resultant of fields originating in matter or in the motions of matter. Therefore, it is, in the light of current knowledge, highly implausible that such a force as f' exists.
>
> Furthermore, this estimate of the implausibility of "F' & G'" reflects *experimental* evidence against "F' & G'", even though this theory has no falsified observational consequences. (Boyd 1973, pp. 7–8)[11]

Boyd's passage is illuminating in two respects. First, it is not only the unobservability of a universal force what makes it bizarre and lacking evidential support. It is also a very implausible concept, in the sense that it is not alike at all to usual forces in another crucial respect: there is nothing in Reichenbach's theorem to let us know about its physical underpinning. Usual forces have a source, for example (typically charges and massive objects); but what is the source of universal forces? Second, the quote underscores that the problematic nature of universal forces is not just a matter of theoretical uneasiness. Universal forces are bizarre not only from the point of view of formal a priori or pragmatic considerations. The difficulties with them are also based on lack of *empirical evidence* to support their reality. Let me clarify this point with yet another quote, this time from a paper by John Norton:

> I must note that the notion of a universal force, as a genuine, physical force, is an extremely odd one. They are constructed in such a way as to make verification of their existence impossible in principle. The appropriate response to them seems to me not to say that we must fix their value by definition. Rather we should just ignore them and for exactly the sorts of reasons that motivated the logical positivists in introducing verificationism. Universal forces seem to me exactly like the fairies at the bottom of my garden. We can never see these fairies when we look for them because they always hide on the other side of the tree. I do not take them seriously exactly because their properties so conveniently conspire to make the fairies undetectable in principle. Similarly I cannot take the genuine physical existence of universal forces seriously. Thus to say that the values of the universal force field must be set by definition has about as much relevance to geometry as saying the colors of the wings of these fairies must be set by definition has to the ecology of my garden." (Norton 1994, p. 165)[12]

---

[11] Kyle Stanford offers a similar account of the matter: "While Eddington, Reichenbach. Schlick and others have famously agreed that general relativity is empirically equivalent to a Newtonian gravitational theory with compensating 'universal forces', the Newtonian variant has never been given a precise mathematical formulation (the talk of universal forces is invariably left as a promissory note), and it is not at all clear that it can be given one (David Malament has made this point to me in conversation). The 'forces' in question would have to act in ways no ordinary forces act (including gravitation) or any forces could act insofar as they bear even a family resemblance to ordinary ones; in the end, such 'forces' are no better than 'phantom effects' and we are left just with another skeptical fantasy. At a minimum, defenders of this example have not done the work needed to show that we are faced with a credible case of non-skeptical empirical equivalence." (Stanford 2001, p. S6, footnote 6)

[12] In his paper Norton introduces this comment only as a sort of side-remark: "As an aside from my main argument . . . " (*ibid.*). His main goal is to disprove the conventionality of geometry, and his main argument is that universal forces finally reduce to "correction-terms" in suitable gauge transformations that preserve the physical meaning of invariants in general covariant formulations of spacetime theories. This implies that the underlying metric in the spacetime theories involved is not affected at all by the introduction of universal forces. As Norton himself

I totally agree with Norton's view regarding universal forces.[13] However, this stance, I think, should not be taken as an ultimate rejection of them as a possible part of scientific theories. Hypotheses are not testable or untestable in a priori terms. For example, new available auxiliary hypotheses could be conjoined to a certain untestable hypothesis and turn it into a testable one. I see no reason why this might not happen in the case of universal forces. That is, *so far as we know*, there is not empirical evidence for the reality of such forces, but future findings might provide good reasons to postulate them in physical theories. A future theory could include observable effects that, at least indirectly, support the reality of a universal force.

It is important to underscore that these remarks hold for universal forces as such, that is, independently of their involvement in Reichenbach's argument. Actually, it seems that this particular argument requires, by definition, that universal forces are not related to any observable effects. The EE between T and T' seems to have as a condition that the universal forces are totally undetectable. However, as I just mentioned – and putting Reichenbach's argument aside –, there might be possible physical theories in which universal forces do relate to observable features. At least this possibility has not been disproven.

The answer to the question of whether Reichenbach's argument involves a challenging case of EE is thus negative. The reason is that, so far as we know, there is no evidential support for the reality of universal forces. Therefore, even though we could concede that Reichenbach's example involves genuine EE and rivalry, this does not mean that we are facing a case of UD, for the theory in which universal forces are absent has more evidence in its favor than its rival. Moreover, the fact that Reichenbach's argument requires that the universal forces involved are totally undetectable suggests that this particular example cannot provide a case of UD, no matter what particular form these forces take within the theory they are a part of. If universal forces are to have any special consequences with respect to EE and UD, it will not be through Reichenbach's example.

---

explicitly acknowledges, this is a refutation of a strong version of the conventionalist thesis. This argument leaves the weak "linguistic-definitional" version untouched – which is Reichenbach's stance – though Norton states that such a version is trivial.

[13]Reichenbach's followers could reply that, since they define force as something which is responsible for a geometrical change, and therefore it essentially depends on the coordinative definitions underlying a physical geometry (see footnote 8 above), then the reality of universal forces is also a matter of convention – and their introduction becomes justified. But this answer only shifts the problem. The usual physical meaning of "force" is much more substantial than a mere stipulation about the presence or absence of geometrical changes. Reichenbach's conventional definition of force is quite debatable.

4 "Total Theories" or "Systems of the World"

The last case of EE in terms of artificial examples I will address is given by "total theories" or "systems of the world". Such theories are defined by providing an account of all possible phenomena, past present and future, in opposition to regular "local" theories that hold for a determinate realm of appearances:

> The thesis of underdetermination of theory choice by evidence is about empirically adequate total science; it is a thesis about what Quine calls 'systems of the world' – theories that comprehensively account for all observations – past, present and future. It is a thesis about theories that entail all and only the true observational conditionals, all the empirical regularities already confirmed by observation and experiment. (Hoefer and Rosenberg 1994, p. 594)

As I mentioned in Sect. 1, Laudan and Leplin introduced an argument intending to show that EE and UD is a surmountable problem in the case of usual local theories. They state that EE between theories is a contingent, time-indexed feature, in the sense that further development of science and technology might break this condition – new available auxiliary hypotheses might lead to diverging predictions, for example. Besides, even if EE remains, the UD of the choice might get broken anyway: if only one of the theories can be encompassed in a more general one, then the evidential support of the latter flows to the encompassed theory but not to its predictively equivalent rival – and thus the evidential tie gets broken.

Hoefer and Rosenberg accept this solution,[14] but they correctly affirm that it cannot work in the case of total theories – that's why they state that the problem of EE and UD is a problem only for total theories. Since Laudan and Leplin's argument makes essential reference to background science, that is, to *other* theories, if we are dealing with systems of the world such other theories are, by definition, not available. All possible auxiliary hypotheses are included in the EE total theories involved, and there cannot be more general theories in which to encompass any system of the world. Therefore, EE in the case of total theories seems to pose a special challenge.

I think that it is true that if a pair of predictively equivalent theories of this kind were given, then the UD involved could not be overcome. However, we do not need to worry about this example of EE either. Even though the very definition of a system of the world precludes that UD could be broken in terms of empirical evidence if EE is given, this definition is problematic in the sense that there is no way for us to know whether a specific theory counts as a system of the world or not.

[14] For a critical reassessment of Laudan and Leplin's argument see (Acuña and Dieks 2013). There we argue that even though their argument does provide a possible way out of the problem, it is not a *guaranteed* solution. The solution that Laudan and Leplin propose essentially depends on the *contingent* development of science, and such a development might not be as required for the solution to be instantiated. New auxiliary hypotheses and new general theories might not be capable of breaking either EE or UD, for example.

There are several ontological and epistemological difficulties with the concept. First, if we are going to take systems of the world seriously, it would have to be shown that the world admits a description by a theory like that. This question involves a metaphysical issue of course: is the set of all natural phenomena regular and coherent enough as to be describable in terms of one single theoretical framework? Second – even if we take for granted that this is possible – is human science capable to provide an alternative, rival, predictively equivalent system of the world? If we discard algorithms and bizarre, parasitic theories this sounds like an extremely unlikely scenario.

It could be argued that the possibility of a total theories-EE scenario has not been disproven, and that this is enough to take the problem seriously. We can concede this, but the problems with the concept of a system of the world do not end here. Recall that the definition involves the property of being empirically adequate for all possible phenomena, past, present and future; but how in the world could we know that a certain (total) theory will be empirically adequate with respect to all future phenomena? Notice that the problem is not that we cannot know whether a certain total theory is true (or empirically adequate) or not; the problem is that since we can never know that a certain theory is empirically adequate with respect to future phenomena implies that we cannot know whether a certain theory is really a system of the world. That is, the very definition of the concept at issue precludes us to know that any candidate-theory is really a total one or not.

Analogously, we cannot know whether a certain theory has all possible phenomena under its scope. It is true that by its form and content a certain theory can claim to be valid in a total way – for all possible phenomena – but the fact that a certain theory intends to be a total one does not necessarily mean that it is. Our world is not like the universe in Poincaré's parable, we cannot accommodate it in a way such that it complies with our theoretical framework. There might always be realms of phenomena that are not accounted for in a theory, even if such a theory intends to be a system of the world. For example, assume that we are facing a case of EE between two total theories. In spite of what the theories say, nothing precludes the possibility that new kinds of phenomena – that have never been observed before and that cannot be accounted for by any of the theories involved – get detected. This already shows that we can never know if the theories involved are total or not. Besides, if such unexpected phenomena are indeed detected, then the problem of EE and UD at issue could be solved à la Laudan and Leplin – the auxiliary hypotheses provided by a new theory that explains the unexpected phenomena could break the predictive equivalence, for example.

The upshot of these remarks for the problem of EE and UD is clear. It is true that if two total theories are EE then the UD of the choice would be a big problem.[15]

[15]If one of the theories includes implausible universal forces, for example, the alternative theory might be better supported by evidence in spite of the EE. That is, the EE between systems of the world would be a big problem granted that both the theories are genuinely scientific and have solid foundations.

However, from the point of view of human scientific knowledge, the very concept of a system of the world is problematic. It is impossible to know whether a certain theory qualifies as a total one. At most, philosophers can speculate about their epistemological and/or metaphysical consequences on a high level of abstraction, but total theories do not present a serious case of EE and UD in the context of the philosophy *of science*. The situation is thus analogous to Descartes' evil-genius argument. It is an interesting and serious issue in metaphysics and general epistemology, but it does not have any particular or relevant consequences for the philosophy of science.[16]

5 Summary and Conclusion

I have considered three examples of artificial examples of EE that have received attention in the philosophy of science literature insofar as they are supposed to imply UD. We have seen that, rightly assessed, none of these examples really entails a problem regarding UD of theory choice. They might be interesting for other reasons – van Fraassen's $TN_{(v)}$ and Reichenbach's argument were originally introduced with a different aim – but they are harmless with respect to the problem that occupies us here. Elsewhere[17] I have argued that neither algorithms nor the Duhem-Quine thesis can be used as sources of problematic EE. This means that the only case where EE and UD can imply a serious problem is in the case of actual scientific theories. However, in scenarios like this Laudan and Leplin's argument offers a possible, contingent way out.[18]

References

Acuña, P., and D. Dieks. 2013. Another look at empirical equivalence and underdetermination of theory choice. *European Journal for Philosophy of Science*. doi: 10.1007/s13194-013-0080-3.

Boyd, R. 1973. Realism, underdetermination, and a causal theory of evidence. *Noûs* 7: 1–12.

[16]Samir Okasha has offered an objection to the cogency of the very concept of a total theory, but along a different line of reasoning. He claims that since the theoretical-observational distinction is not absolute, but context-dependent – a certain term in a theory counts as theoretical, but the same term in a different theory can count as observational – neither the observational content nor the theoretical apparatus of a system of the world can be defined: "If we are even to understand this suggestion [that EE between two total theories leads to UD], let alone endorse it, we must have a criterion for deciding which side of the divide an arbitrarily chosen statement falls on. But such a criterion is precisely what the minimal, context-relative theory/data distinction does not give us. If that distinction is all we have to go on, we can get no grip on what it means for our 'global theory' to be underdetermined by the 'empirical data', nor indeed on what a 'global theory' is even supposed to be." (Okasha 2002, p. 318)

[17](Acuña and Dieks 2013).

[18]See (Laudan and Leplin 1991).

Earman, J. 1970. Who's afraid of absolute space? *Australasian Journal of Philosophy* 48: 287–319.
Havas, P. 1964. Four-dimensional formulations of Newtonian mechanics and their relation to the special and the general theory of relativity. *Reviews of Modern Physics* 36: 938–965.
Hoefer, C., and A. Rosenberg. 1994. Empirical equivalence, underdetermination, and systems of the world. *Philosophy of Science* 61: 592–607.
Laudan, L., and J. Leplin. 1991. Empirical equivalence and underdetermination. *The Journal of Philosophy* 88: 449–472.
Norton, J. 1994. Why geometry is not conventional: The verdict of covariant principles. In *Semantical aspects of spacetime theories*, ed. U. Majer and H.-J. Schmidt, 159–167. Mein-heim/Leipzig/Wien/Zurich: Wissenschaftsverlag.
Okasha, S. 2002. Underdetermination, holism, and the theory/data distinction. *The Philosophical Quarterly* 52: 303–319.
Poincaré, H. 1952 [1901]. *Science and hypothesis*. New York: Dover.
Reichenbach, H. 1958 [1928]. *The philosophy of space & time*. New York: Dover.
Sklar, L. 1974. *Space, time, and spacetime*. Berkeley/Los Angeles/London: University of California Press.
Stanford, K. 2001. Refusing the devil's bargain: What kind of underdetermination should we take seriously? *Philosophy of Science* 68 (supplement: Proceedings of the 2000 biennial meeting of the Philosophy of Science Association. Part 1: contributed papers): S1–S12.
Stein, H. 1970. Newtonian spacetime. In *The Annus mirabilis of Sir Isaac Newton*, ed. R. Palter, 258–284. Cambridge, MA: MIT Press.
Van Fraassen, B. 1980. *The scientific image*. Oxford: Clarendon.

The Measurement Problem Is Your Problem Too

Ronnie Hermens

1 Introduction

Since Worall (1989) introduced structural realism (SR) as being "the best of both worlds" there is a growing support for the position in the philosophy of physics. In fact, support is drawn from contemporary physics for ontic structural realism (OntSR)[1]: the view that the structures discovered by fundamental science are the primitive metaphysical building blocks. A natural intuition behind this arises from the fact that the abstract formulation of much of contemporary physics has proven to be hard to unify with traditional metaphysical views. It then seems natural to ground the concerning physical theories on an equally abstract metaphysical theory; one that takes structures as the primitive building blocks instead of objects.

But of course more than just an intuition is expected for a sound argument for OntSR. Moreover, for any such argument one should take into account that one cannot draw metaphysical conclusions from science without any metaphysical presuppositions. In this paper I am concerned with these presuppositions when it comes to the standard argument for OntSR from quantum mechanics (QM). In particular, I focus on the question of whether foundational problems for QM pose problems for the argument for OntSR.

In Sect. 2, I first repeat the by now classical argument for OntSR drawn from the difficulties with the principle of identical indiscernibles in QM. Some technical details concerning this argument are further evaluated in Sect. 3. In Sect. 4, I argue

Imitation is the sincerest form of flattery cf. Hájek (2007).

[1]Cf. Lam and Esfeld (2012) and references therein.

R. Hermens (✉)
Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52,
9712 GL Groningen, The Netherlands
e-mail: r.hermens@rug.nl

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_32,

why supporters of this argument (or any other argument from QM) should be worried about the foundational problems for QM, and the measurement problem in particular. In Sect. 5 it is argued that in the light of possible solutions to the measurement problem the support for OntSR becomes questionable. Finally, Sect. 6 evaluates the possibility of saving the support for OntSR by declaring the measurement problem irrelevant for this point. I conclude that this support can only be regained by presupposing elements of OntSR that are not likely to be accepted by anyone not already convinced by this metaphysical position. Strongly put, there is no real support for OntSR from quantum mechanics after all.

2 The Standard Argument

The standard argument from QM in favor of OntSR involves Leibniz's principle of the identity of indiscernibles (PII).[2] I first give a short exposition of this argument before delving into some of the details later on. The metaphysical principle states that, necessarily, if two objects are indiscernible, then they are identical:

$$\Box(\forall a \forall b : \neg \mathrm{Disc}(a, b) \rightarrow a = b). \tag{1}$$

Or in other words, discernibility is necessary for identity.

Much of the recent debate on this principle has been on the question whether or not it can be maintained for particles in QM.[3] This issue of course relies on what is allowed to discern between two objects. The most common notion states that two objects are indiscernible if they share all their properties, i.e., for any P: $P(a) \leftrightarrow P(b)$.

On this reading quantum mechanical particles are indiscernible unless identity is allowed to occur as a (primitive) property. For example, one could then just take P to be "is identical to a" which would then discern a from b. This latter option short circuits the entire discussion on PII and requires the adoption of some thick metaphysical notion like haecceity or primitive thisness. On rejection of such notions, it is then often suggested that quantum particles are not individuals. This may be contrasted to classical mechanics in which particles are often considered to be discernible by their spatiotemporal properties. Thus for classical mechanics PII suggests a metaphysics of particles while for QM it does not.

The lesson drawn from these observations is roughly the following. For QM the individuality of particles does not follow from PII unless it is added as a

[2] Cf. Ladyman (1998), French and Ladyman (2003), Ladyman and Ross (2007), and Muller (2011).

[3] Cf. Saunders (2006), Dieks and Versteegh (2008), Muller and Saunders (2008), Muller and Seevinck (2009), and Dieks and Lubberdink (2011).

surplus. The individuality of particles is then underdetermined by QM[4] and thus poses a dilemma for the object-metaphysician. OntSR dissolves this dilemma by the argument that, since QM does not require the individuality of particles, one may reject their metaphysical significance altogether:

> [...] the way particle permutations are treated in quantum mechanics is quite distinct from classical physics and the above metaphysical underdetermination arises. Thus, we have a metaphysical package of (individual) objects to which the mathematical and physics of quantum theory is applied and which undermines that very package we started with. From this perspective, the objects play only a kind of heuristic role (see French 1999), allowing us to apply the (classical) mathematics and hence getting us up to the (group-theoretical) structures as it were, but once we're there the objects can be dispensed with. (French and Ladyman 2003, pp. 41–42)

And it is phrased more clearly by one of the opponents of OntSR:

> In the face of the underdetermination between individuality and non-individuality, then, it seems [...] more plausible to dispense with the traditional 'individual entities' ontology and adopt [OntSR]. Since the latest developments in microphysics strongly suggest that the very concept of individual entity is a surplus, and relations and properties expressed in structural terms are all we need in order to construct an exhaustive picture of the world, it makes perfect sense to say that there actually isn't anything else. (Morganti 2004, p. 88)

This concludes the quantum mechanical motivation for OntSR. Some issues evolving around this argument will be the topic of the next section. This will also form the prelude to the question to be discussed in Sect. 4, namely, why should a motivation for OntSR from QM be taken serious in the first place? That is, the underdetermination w.r.t. PII in QM is a property of the theory QM, and at first glance it seems that OntSR is only argued for as a metaphysical stance within this theory. Clearly more is intended, and the question arises of how the motivation for OntSR stretches beyond this theory.

3 Discernibility and the Quantum Formalism

When only considering properties that don't presuppose individuality, the failure of the principle is not restricted to quantum mechanics. Perhaps the most famous counterexample is a universe consisting solely of two identical spheres separated by

[4]The use of the term "underdetermination" in this context is somewhat controversial in philosophy of science. On the other hand, it is also commonplace within the present topic. For a defense of the use of this term I refer to the authority of van Fraassen (1991, p. 481): "The phenomena underdetermine the theory. There are in principle alternative developments of science, branching off from ours at every point in history with equal adequacy as models of the phenomena. Only angels could know these alternative sciences, though sometimes we dimly perceive their possibility. The theory in turn underdetermines the interpretation. Each scientific theory, caught in the amber at one definite historical stage of development and formalization, admits many different tenable interpretations. What is the world depicted by science? That is exactly the question we answer with an interpretation and the answer is not unique."

a distance of 2 miles as introduced by Black (1952). In such a universe any property that applies to one sphere also applies to the other. Even spatiotemporal properties can only be invoked as discerning properties if one presupposes a fixed reference point or background coordinate system for describing the positions of the spheres. However, this boils down to presupposing a primitive thisness for space-time points irrespective of the objects that may inhabit them. And a similar presupposition can be required for symmetric configurations of particles in classical mechanics (after all, Black's universe may be considered just a particular instantiation of a classical mechanical world).

But as noted earlier, the failure of PII may depend on the notion of discernibility used. One could argue that any reading may be satisfactory as long as it doesn't presuppose individuality (i.e., doesn't add the metaphysics to the theory but rather extracts it from it). A weaker sufficient condition for discernibility that does discern Black's spheres is the existence of an irreflexive symmetric relation R on the objects:

$$\forall a \forall b : \text{ for some } R : \neg R(a,a) \wedge R(a,b) \rightarrow \text{Disc}(a,b). \tag{2}$$

Indeed, the relation "is at a distance of two miles from" is irreflexive and symmetric.

It has been shown that with this notion of weak discernibility, PII actually also holds for QM.[5] This indicates that quantum particles may be seen to be individuals after all. On the other hand, it has been argued that, for QM, weak discernibility is not as suggestive for identity as for the classical symmetric spheres (Dieks and Versteegh 2008; Dieks and Lubberdink 2011). To see why this may be so, consider the quantum mechanical analogue to the two spheres of two fermions in the singlet state

$$|\psi\rangle = \frac{1}{2}\sqrt{2}\,(|\uparrow\rangle_1 |\downarrow\rangle_2 + |\downarrow\rangle_1 |\uparrow\rangle_2). \tag{3}$$

The irreflexive relation that weakly discerns them is given by "has spin opposite to". According to Dieks and Lubberdink (2011, p. 1062), the arguments for individuality in QM "hinge on a silent premiss, namely that the indices not only play a mathematical role but also possess *physical* significance." They go on to argue that for Black's spheres this poses no problem: "our mind's eye sees these spheres at different distances or in different directions before us; in thought we break the symmetry" (*ibid*., p. 1063). This is to be contrasted with the quantum case where "relations have a standard interpretation not in terms of what is actual, but rather via what *could happen in case of a measurement*" (Dieks and Versteegh 2008, p. 934).

I believe the argument of Dieks et al. has quite some appeal. The state description of the two spheres refers directly to something in the world which makes the indices

[5] Cf., Muller and Saunders (2008), Muller and Seevinck (2009), and Caulton (2013).

physically significant. The quantum state description, on the other hand, only has an indirect physical significance. At least, on the standard (Copenhagen) interpretation it does:

> The entire formalism is to be considered as a tool for deriving predictions, of definite or statistical character, as regards information obtainable under experimental conditions described in classical terms and specified by means of parameters entering into the algebraic or differential equations of which the matrices or the wavefunctions, respectively, are solutions. These symbols themselves, as is indicated already by the use of imaginary numbers, are not susceptible to pictorial interpretation. (Bohr 1948, p. 134)

And on such a reading it seems appropriate that indices in the state description have no physical significance.

However, it is not clear why one should be committed to this particular reading of the formalism. In fact, it would seem that if one wishes to draw metaphysical conclusions from QM, to a certain extent these conclusions should not hinge on a particular reading of the formalism. What then to make of this weak discernibility w.r.t. the motivation for OntSR from QM? At this point Muller (2011) provides an answer. He emphasizes that, while quantum particles are weakly discernible (by relations), they are not absolutely discernible (by properties). Such objects are then called relationals (reserving the term 'individuals' for absolutely discernible objects). These observations amount to the conclusion that

> [s]tructuralist objects, if they must exist, *should be* relationals. Well, we have seen that is exactly what elementary particles demonstrably *are*: relationals. This leads us to OntSR. (*ibid.*, p. 231)

This then is an argument independent of how one views the wave function. What it leaves outside of its scope, however, are interpretations in which the wave function is not a complete state description and additional properties of systems are added. As seen in the previous section, one of those properties could be haecceities. Scientifically more interesting however is the case of Bohmian mechanics, where locations are added as properties for the particles. On this interpretation (leaving Blackian symmetrical configurations aside), quantum particles are absolutely discernible. How does the argument for OntSR fare with this option on the table? This is a question I will discuss in Sect. 5 after the evaluation of another question that has been lying in the background: to what extent should we take metaphysical constraints from quantum mechanics serious outside the scope of this theory? That is, the argument for OntSR is supposed to work as a general argument for this position within metaphysics, not just for the foundations of quantum mechanics. And how this generality is achieved is a subtle matter.

4 Quantum Mechanics as Fundamental Physics

What warrants the role of QM in the PII-argument? To contrast, classical mechanics does suggest the individuality of particles, and may be taken to argue in favor of a metaphysics of things and against OntSR. Why does QM outweigh this argument?

The answer found in Ladyman and Ross (2007) is that QM is a fundamental theory. This idea is reflected in the quote from Morganti (2004) in Sect. 2: it is because QM constructs "an exhaustive picture of the world" that we can use it as a guide for metaphysical lessons. This is captured more explicitly in Ladyman's and Ross's primacy of physics constraint (PPC) which states:

> Special science hypotheses that conflict with fundamental physics, or such consensus as there is in fundamental physics, should be rejected for that reason alone. Fundamental physical hypotheses are not symmetrically hostage to the conclusions of the special sciences. (Ladyman and Ross 2007, p. 44)

The intuitive idea behind this is that fundamental physics is more in the business of discovering real structures that will persist throughout the development of science than the special sciences are. Then the argument for OntSR from QM has bite because QM is part of fundamental physics.

Intuitively it seems natural to count QM as a fundamental theory. One may even require that counting QM as fundamental may serve as a test for whether any definition of "fundamental" is appropriate. On the scientific level this seems a fine approach because not much hinges on the idea of fundamentality. However, on the current metaphysical level the fundamentality of a scientific theory is decisive for its metaphysical impact. And to avoid circularity one requires a notion of fundamentality that doesn't presuppose metaphysical relevance. A further inquiry is thus desirable.

First one may note that QM is not fundamental in the intuitive sense issued earlier; strictly speaking, QM does not construct an exhaustive picture of the world. Indeed, many things are left out of the picture such as gravitation. Obviously this take on "fundamental" is too restrictive since it would only apply to a theory of everything. No scientific theory available at the moment constructs an exhaustive picture.

A more modest criterion for fundamentality is found in *ibid*. where a definition of a special (i.e., non-fundamental) science is given.

> [A] science is special iff it aims at generalizations such that measurements taken only from restricted areas of the universe, and/or at restricted scales are potential sources of confirmation and/or falsification of those generalizations. (*ibid*., p. 195)

This does seem a more appropriate reading of fundamentality, but also shows some tension with QM. On traditional views of QM, application of the theory requires a sharp division between the system under investigation and the measurement setup used. Only the former is assigned a quantum state, while the latter receives a classical description. In particular, traditionally the idea of a quantum state of the universe is considered troublesome. Thus at least one scale does not obviously provide a potential source for confirmation or falsification. Hence, on the adopted reading of "special science" and the orthodox reading of QM, QM is a special science, i.e., not fundamental.

It seems natural to argue that this concerns a peculiarity of the *formulation* of QM rather than a lack of fundamentality of the theory. However, to solve this peculiarity by a reformulation of the axioms of the theory is precisely to engage in solving

the infamous measurement problem. Attempts to solve this problem have spawned a number of distinct interpretations of QM which draw quite distinct ontological pictures of the way the world is. If support for OntSR is to be grounded, it has to be shown that this support is unaffected by the plurality of pictures. This may be done by arguing that (1) the argument for OntSR is independent of this plurality, or (2) that support for OntSR holds for all possible pictures, or (3) that any picture that does not support OntSR is unsatisfactory.

Discussion of the first option is postponed to Sect. 6, while the options (2) and (3) may be investigated in tandem. As seen at the end of Sect. 3, when it comes to interpretations that leave the state formalism of QM intact, Muller's argument may provide a type (2) response. However, for hidden-variable solutions to the measurement problem the evaluation of PII may lead to a different outcome. So the problem of support for OntSR lies with these solutions in particular, and whether option (3) applies here is the topic of the next section.

5 Hidden Variables and Naturalized Metaphysics

The main ingredient for any hidden-variables theory is that the quantum state does not provide a complete description of a system. There are additional (metaphysical) properties not encoded by the wave function that play an essential role in bringing about measurement outcomes. The paradigm example is of course Bohmian mechanics in which the wave function is supplemented with the positions of all particles to obtain a complete state description. As such, Bohmian mechanics is similar to classical mechanics in that it makes particles individuals by virtue of their position properties.

In dismissing this option as a viable solution to the measurement problem, Ladyman and Ross (*ibid.*) put their naturalized view in metaphysics to work. Like the possibility of haecceity, the option of hidden-variables is deemed "scientifically unmotivated and ad hoc, and only motivated by philosophical rather than physical considerations" (p. 181). As such, it is not viewed as part of fundamental physics and by PPC it need not be taken seriously metaphysically.

The ease with which hidden-variable approaches are dismissed appears as somewhat surprising to me. The completeness of the wave function has played an important role in the foundations of quantum mechanics since its founding days. To dismiss the EPR-argument as "unscientific", for me, requires more motivation than given by Ladyman and Ross (2007). For one, one would like an argument showing that any additional ontology in a hidden-variable approach to the measurement problem is in fact a surplus, that only satisfies philosophical needs. This in turn requires showing that there are viable solutions to the measurement problem that do not require the surplus *and* that the surplus is indeed non-scientific.

But as it stands the prospects are grim on both accounts. Irrespective of whether Bohmian mechanics (or any other hidden-variables approach for that matter) actually provides a satisfactory solution to the measurement problem, it should be

noted that other options are also not faring too well. Like Bohmian mechanics, collapse theories like GRW face difficulties with the requirement of Lorentz-invariance for further generalizations. Everettian approaches on the other hand face the familiar problems of the preferred basis and of probabilities. And other problems arise for other attempted solutions. That is, in the literature on the measurement problem it is not so much a discussion of *which* interpretation provides the correct solution, but rather *if* any interpretation actually provides a solution.

On the second issue one may note that a hidden-variable theory is non-scientific to the extent that the hidden variables are actually hidden. However, that they are hidden is not a requirement on this approach.[6] In fact, possible discrepancies with QM are to be embraced and actually are a source for (admittedly tentative) scientific research. An example of this for Bohmian mechanics is Valentini (2010). At any rate, it seems premature to deem a possible solution to the measurement problem unscientific merely because it posits an extension to the standard formalism. And if this extension suggests the identity of particles this should be taken seriously by a naturalized metaphysician. And to emphasize, this does of course not discredit OntSR as a viable metaphysical position, but merely the idea that modern physics necessitates this position.

6 The Measurement Problem Is Your Problem Too

In Sect. 4 I argued that, in order for QM to be a fundamental theory, a solution to the measurement problem is required. And as argued in the previous section investigations of this problem allow for options that are suggestive towards object metaphysics and that are to be taken seriously scientifically. To maintain that quantum mechanics is still suggestive towards OntSR one could then argue that the measurement problem itself is actually a non-issue for this point.

One way to motivate this view is to object that the classification of QM as a special science hinges on the particular notion of fundamentality that is adopted. That is, the particular notion of fundamentality requires a certain level of physical precision of a theory in order to check whether it applies or not.[7] A requirement not everyone may endorse. And although in orthodox readings of QM the theory does not apply to the totality of everything, it does apply to anything piecewise. And

[6]The term "hidden variables" has some unfortunate connotations. First of all, the additional variables may be empirically accessible. In fact, in Bohmian mechanics the positions are the variables whose values are revealed by measurements (see also the defense of Bohmian mechanics given by Bell 1982 (reprinted in (Bell 1987))). What is intended here, however, is the question of whether a hidden-variables approach should be entirely empirically equivalent to orthodox QM. And in this sense the additional variables need not stay hidden either.

[7]For one, it requires an explication of measurement processes and their outcomes as physical processes. A requirement strongly advocated by Bell (1990).

this may be considered sufficient for fundamentality.[8] The split between system and apparatus that underlies the measurement problem may then just be taken for granted. This is the attitude adopted by Ladyman and Ross (2007) who seem to accept the measurement problem as a trait of the theory:

> [T]he application of the quantum formalism to macroscopic objects is not necessarily justified [...] [T]he representation of macroscopic objects using quantum states can only be justified on the basis of its explanatory and predictive power and it has neither. [...] The predictive success of QM in this context consists in the successful application of the Born rule, and that is bought at the cost of a pragmatic splitting of the world into system and apparatus. (*ibid.*, p. 182)

Now it is relatively uncontroversial that direct application of the QM formalism to macroscopic objects is problematic or even nonsensical. But to use this as a defense for leaving it unresolved in the current metaphysical discussion is to not fully appreciate the issue underlying the measurement problem. That is, that without a solution to the measurement problem, the quantum formalism is in sheer conflict with ordinary descriptions of macroscopic objects. It is uncontroversial that pragmatically the splitting is necessary. The problem is however that, without some additional ontology in the theory, the splitting is also necessary in principle. Without a solution there is no clear complete picture of the ontology of the world nor of the precise constraints posed by QM.

Now one may ask if these requirements really need to be met in order for the PII-argument to work. Is it really necessary to have a set of full answers to what the world is like if QM is correct? It seems not everyone would concede to give a "yes" here. In fact, Ladyman time and again (1998, 2007, 2009) adheres to the motto that

> imaginability must not be made the test for ontology. The realist claim is that the scientist is discovering the structures of the world; it is not required in addition that these structures be imaginable in the categories of the macroworld. (McMullin 1984, p. 14)

And when applying this motto to the PII-argument one arrives at the conclusion that an answer to what the quantum state represents is not required for deriving conclusions from the state formalism.

It seems to me that this attitude is indeed what it takes to make the PII-argument for OntSR work. However, I also think that this attitude presupposes enough traits of OntSR to make the argument very weak. Imaginability seems a perfectly natural demand within naturalized metaphysics (though not a necessary one). And those who take imaginability seriously have reason to take the measurement problem seriously and, consequently, also the possibility of an object metaphysics. This is not to say that the PII-discussion is completely vacuous of course. It still shows that modern physics is not unproblematically compatible with classical metaphysical intuitions. But I won't go as far as Ladyman and Ross (2007) to state that every thing *must* go. Instead, I would settle for the claim that every thing *may* go.

[8] I took this idea from Timpson (2008, p. 590) who posed it as part of a possible defense for Quantum Bayesianism.

Acknowledgements The author would like to thank the following people for constructive comments and discussions: D. Dieks, F.A. Muller, J.W. Romeijn, M.P. Seevinck, and two anonymous referees. This work was supported by the NWO (Vidi project nr. 016.114.354).

References

Bell, J.S. 1982. On the impossible pilot wave. *Foundations of Physics* 12(10): 989–999. Reprinted in Bell 1987.

Bell, J.S. 1987. *Speakable and unspeakable in quantum mechanics*. Cambridge: Cambridge University Press.

Bell, J.S. 1990. Against 'measurement'. *Physics World* 3(8): 33–40.

Black, M. 1952. The identity of indiscernibles. *Mind* 61(242): 153–164.

Bohr, N. 1948. On the notions of causality and complementarity. *Dialectica* 2(3–4): 312–319.

Caulton, A. 2013. Discerning "indistinguishable" quantum systems. *Philosophy of Science* 80(1): 49–72.

Dieks, D., and A. Lubberdink. 2011. How classical particles emerge from the quantum world. *Foundations of Physics* 41: 1051–1064.

Dieks, D., and M.A.M. Versteegh. 2008. Identical quantum particles and weak discernibility. *Foundations of Physics* 38: 923–934.

French, S. 1999. Models and mathematics in physics: The role of group theory. In *From physics to philosophy*, ed. J. Butterfield and C. Pagonis, 187–207. Cambridge: Cambridge University Press.

French, S., and J. Ladyman. 2003. Remodelling structural realism: Quantum physics and the metaphysics of structure. *Synthese* 136(1): 31–56.

Hájek, A. 2007. The reference class problem is your problem too. *Synthese* 156: 563–585.

Ladyman, J. 1998. What is structural realism? *Studies in History and Philosophy of Science* 29(3): 409–424.

Ladyman, J. 2009. Structural Realism. In *The Stanford encyclopedia of philosophy*, ed. Edward N. Zalta. http://plato.stanford.edu/archives/sum2009/entries/structural-realism/

Ladyman, J., and D. Ross. 2007. *Every thing must go*. Oxford: Oxford University Press.

Lam, V., and M. Esfeld. 2012. The structural metaphysics of quantum theory and general relativity. *Journal for General Philosophy of Science* 42(2): 243–258.

McMullin, E. 1984. A case for scientific realism. In *Scientific realism*, ed. Jarrett Leplin, 8–40. Berkely: University of California Press.

Morganti, M. 2004. On the preferability of epistemic structural realism. *Synthese* 142: 81–107.

Muller, F.A. 2011. Withering away, weakly. *Synthese* 180: 223–233.

Muller, F.A., and S. Saunders. 2008. Discerning fermions. *British Journal for the Philosophy of Science* 59: 499–548.

Muller, F.A., and M.P. Seevinck. 2009. Discerning elementary particles. *Philosophy of Science* 76(2): 179–200.

Saunders, S. 2006. Are quantum particles objects? *Analysis* 66(1): 52–63.

Timpson, C.G. 2008. Quantum Bayesianism: A study. *Studies in History and Philosophy of Modern Physics* 39: 579–609.

Valentini, A. 2010. Inflationary cosmology as a probe of primordial quantum mechanics. *Physical Review D* 82(6): 063513(1–22).

van Fraassen, B. 1991. *Quantum mechanics: An empiricist view*. Oxford: Oxford University Press.

Worall, J. 1989. Structural realism: The best of both worlds? *Dialectica* 34(1–2): 99–124.

Pros and Cons of Physics in Logics

Petr Švarný

1 Introduction

Physics relies on formal systems[1] as its backbone that allows it to present its ideas in a clear manner. Still, it maintains a significant need of synthetic knowledge, it demands testing of its results. On the other hand, logics can use physics as a source of inspiration for new approaches to reasoning. As opposed to physics, logics are analytical and even if some part of physics they have taken inspiration from would be proven as not being in accordance with the world, then that given logic could live on.

We investigate the following two approaches. We start out from logics with some well known logical system and we try to formalize physics with it. The benefit of this approach is that we know the tool very well and can focus only on modelling the desired properties of the physical system. The second approach is to have a physical property for which we try to find a logical framework and if needed, we define it anew. These approaches give rise to at least two questions: Do physics add fundamentally new ideas or approaches to logics? On the other hand, do logics allow physicists to understand better the chosen physical problem than they would do without them? These questions might remind the reader of one

I must give much credit and thanks to Dennis Dieks for his support. Výstup projektu Vnitøních grantù 2012 Filozofické fakulty UK. Made with the support of the Internal grant of the Faculty of Arts, Charles University, 2012.

[1] A formal system being "abstract, theoretical organization of terms and implicit relationships that is used as a tool for the analysis of the concept of deduction". (Encyclopædia Britannica Online 2013)

P. Švarný (✉)
Department of Logic, Charles University in Prague, Prague, Czech Republic
e-mail: svarnypetr@gmail.com

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_33,

of the classical debates, namely the one about the frontiers between synthetic and analytic propositions. This debate has its proper place in logics as it was investigated not only by Kripke, but the whole field of paraconsistent logics presents a work on this subject (Lacey 1996). This article, however, does not address this classical debate.

We introduce both of these approaches in more detail in the following two sections. We show at least one example of the given approach in each section.[2] A general and rather brief comment about the relation of physics and logics is presented in the last section, followed by some suggestions for further research.

2 Physics in Logics

Although some people might think that classical logic is too simple to capture a physical theory, the opposite is proven by the Hungarian group around prof. Németi (Andréka et al. 2011).

They axiomatize the theory of special relativity in first-order logic (FOL) and thereby strive to find a minimal number of convincing axioms that could be used to derive all the predictions given by relativity in physics. The goal is to clarify the axioms, their relation, and their influence, thereby also eliminating any tacit assumptions connected to the theory of relativity.

The benefits, as given by the Németi group to explain the choice of FOL as a basis, are its complete inference system, unambiguous syntax and semantics, and that it is a fragment of natural language. Thanks to these properties, axioms should avoid tacit assumptions and any rules of thumb from physics.

3 Logics Based on Ideas from Physics

Approaches that take the opposite direction focus on some physical properties and then find a formal model that investigates the behaviour of truth under these specific conditions. First, let us remark that physics themselves clearly are a formal system already. However, the logical and physical formal systems serve both different purposes. The logical system isn't meant for modelling physical laws of the system but mainly to model the behaviour of truth in the given environment.[3]

This inspiration from physics can have two possible forms. The first one takes a general physical property, creates models based on this property, and thereafter investigates the behaviour of truth in these models to reason about the physical system itself. The outcome is different in the second approach, where the end result

[2] Due to limited space, we confine detailed introductions to given logics to the cited papers.

[3] Usually the truth of a sentence given by the specific language of that physical environment.

is not about investigating the original physical system but it is an attempt to present a new approach to the treatment of truth. This new representation of truth only takes into account also the original specific physical limitations.

3.1 Branching Models

Branching models, as we call in general all the work related to Belnap's Branching space-time models (BST), can be taken as physically motivated formal systems. This follows from the original problem that they aim to solve which is "*How can we combine relativity and indeterminism in a rigorous theory?*" (Belnap 1992). From the point of view of logic, relativity itself is not a relevant problem to investigate. In addition, BST models also try to phrase their results as close to the physical laws as possible, this can be seen on multiple attempts to capture and explain physical phenomena in these models (e.g. Müller 2002; Placek 2009; Placek and Belnap 2011).

3.2 Quantum Logic

The second approach can be illustrated on the case of quantum logic (QL). Already the founding fathers of QL, Birkhoff and Neumann, noted that "One of the aspects of quantum theory which has attracted the most general attention, is the novelty of the logical notions which it presupposes . . . " (Dalla Chiara and Giuntini 2002). The idea of quantum states and their superpositions is taken from physics and allows a new treatment of truth values. There is a multitude of approaches how to capture quantum properties with logics but let us take orthologic (OL) or orthomodular quantum logic (OQL) as the main two possibilities.[4] Both can serve as an example for our case as opposed to articles like (Blute et al. 2003), which investigates also logics in the context of quantum physics, but uses linear logics to model quantum states and hence lacks the same level of inspiration from physics ad OL or OQL do.

4 Paths Not Taken

As we have seen, there is a difference in motivation and technique between the two main points of view. Although motivations from both camps can be varied, there are two main questions that can be asked. A logician could ask whether some aspect

[4]OQL being presented for example in Hardegree (1981), where it is also shown how logic, mathematics, and physics can be interconnected. This is visible at the point, where the author explains that orthomodular lattices are chosen over Hilbert lattices because of the mathematical properties of Hilbert spaces. However, the resulting logic is an attempt to formalize quantum computation.

of how truth behaves in our day-to-day experience is omitted by logicians, but has some property similar to a physical system. A question coming from the physics camp could be, whether there are some physical systems and properties that are not formalized yet, but could be. Both cases lead to either a physics inspired logical system or the application of logics in physics. Realizing that there is this relation allows us to examine such situations on purpose and profit from them.

4.1 Branching Axiomatizations

As our first example, we mention once again branching models. They were primarily meant to capture physical properties, hence there was less focus on their logical aspects. The usual logical questions, such as completeness or axiomatization, were not addressed. As shown in Švarný (2013), axiomatization of some of the properties demanded by the models of Branching continuations is a very complicated task and asks for the use of more than just modal temporal logics. With the use of hybrid temporal propositional formulae it is already possible to capture more properties of Branching continuations. These limits, however, are clearly given by common limits of logic as for example the accessibility relation in Branching continuations should have the antisymmetry property, but this cannot be captured by any modal formula. The language of Branching continuations does not consist of sufficient tools to capture this property. Its language would need something similar to nominals as they are known in hybrid logic. Similarly, also other properties of the Branching continuations models fail to be captured by any formula in the language specified for these models. Hence the axiomatization of branching structures remains an open question for logicians. If a suitable language and axioms would be found for branching models, it would allow a similar investigation to that made by Andréka et al. (2011).

4.2 Epistemic or Doxastic Vector Logic

Epistemic logics focus on the subject of knowledge and there is already a great range of possibilities how to formalize it. However, if one thinks about opinions rather than knowledge, then sometimes words like momentum, weight, or impact. These words come from classical mechanics and it could actually serve us as a source of inspiration. We can sketch a logical system that investigates doxastic states of agents while using similar terminology and techniques as in simple classical mechanics. Thanks to this system, we can formalize opinions and how they can sway based on other agents' actions.[5]

[5] At least a note needs to be dedicated to the fact that there exists something called Vector logic. However, Vector logic is not inspired by physics and represents an interesting case for itself. "*Vector logic is a matrix-vector representation of the logical calculus inspired in neural network*

The idea of vector like logic already appeared in the form of arrow logic (Venema 1996). This logic presents a framework for any transition based logic and one of its motivations are transitions amongst opinions or a way how to formalize presuppositions.

We mentioned that this should be a logic inspired by physics. Arrow logic itself already includes identity arrows, related arrows or composed arrows. What could be taken from physics to further arrow logic are different contexts of use, in physics environments of propagation. An arrow, representing some knowledge, could be subject to effects similar to refraction.

Vector logic capturing opinions could serve as an alternative to classical dynamic epistemic logics or their evidence counterparts, and serve as a tool for modelling decision problems or arguments with agents.

5 Summary

We investigated three possible ways how physics and logics can relate to each other. The first option is to look at physics by using a well known logic trying to model a physical theory with it. The second and third options take the opposite direction. Either start out from a physical theory and try to formulate some new logic based on that theory or use the physical mechanism itself as a new logic for some different purpose. We have shown examples of these approaches and expressed the view that acknowledging these three correlations allows us to find new logics or formalizations of physics.

The main aim of this paper was to present a starting point for a deeper investigation of the influence pathways between logics and physics. We showed that the differences between the studied fields present a challenge in some cases, they nevertheless allow fruitful interaction between both of them. The problem that still stays unresolved is the centuries old tension between analytic and synthetic truths. If we have a logic based on a physical theory and it tries to find all the statements that are entailed by this physical theory, what does it say about physics when the logical system is inconsistent, isn't complete, or has some other specific property? Not much, because we chose the logical system on the beginning when it seemed to be the most suitable for the physical model. Logic remains a formal tool that tells us what a system of rules entails and what reasoning is valid in it. It, however, does not claim anything about the physical world itself. Nevertheless, logic can be a powerful tool for explanation and clarification even in physics.

On the other hand, does physics shed any light on logics? Yes and it sheds surprisingly more light than one would expect. Physics can contribute to logics in a

models" (Mizraji 2008). In other words, it is more a biology based inspiration (human brain biology). Hence, it is a logic, i.e. study of valid reasoning, based on the actual process of thinking as it is happening in our brain.

different way than mathematics do. Physics bring new ideas, as they do not have to follow common rules of reasoning necessarily and can come up with surprising situations that challenge logics and allow them to expand into new territories. Nevertheless, as soon as a new logic is born, it can live its life freely without remembering its physical roots. Still it remains a question to what extent such logics remain interesting, for physicists or logicians, if their source of inspiration becomes revoked due to some physical experiment.

References

Andréka, H., J. Madarász, I. Németi, and G. Székely. 2011. Logical analysis of relativity theories. *Hungarian Philosophical Review* 2011: 204–222.

Belnap, N. 1992. Branching space-time. *Synthese* 92: 385–434.

Blute, R.F., I.T. Ivanov, and P. Panangaden. 2003. Discrete quantum causal dynamics. *International Journal of Theoretical Physics* 42(9): 2025–2041.

Dalla Chiara, M.L., and R. Giuntini. 2002. Quantum logics. In *Handbook of philosophical logic*, 129–228. Amsterdam: Springer.

Encyclopædia Britannica Online. 2013. Formal system. Retrieved from http://www.britannica.com/EBchecked/topic/213751/formal-system

Hardegree, G.M. 1981. An axiom system for orthomodular quantum logic. *Studia Logica* 40(1): 1–12.

Lacey, A.R. 1996. *A dictionary of philosophy*. London: Psychology Press.

Mizraji, E. 2008. Vector logic: A natural algebraic representation of the fundamental logical gates. *Journal of Logic and Computation* 18(1): 97–121.

Müller, T. 2002. Branching space-time, modal logic and the counterfactual conditional. In *Non-locality and modality*, 273–291. Amsterdam: Springer.

Placek, T. 2009. Possibilities without possible worlds/histories. *Journal of Philosophical Logic* 40: 1–29.

Placek, T., and N. Belnap. 2011. Indeterminism is a modal notion: Branching spacetimes and Earman's pruning. *Synthese* 187: 1–29.

Švarný, P. 2013. Wally axiomatics of branching continuations. *Acta Universitatis Carolinae – Philosophica et Historica* 2/2010: 75–86.

Venema, Y. 1996. A crash course in arrow logic. In *Arrow logic and multimodal logic*. CSLI, Stanford.

How Fundamental Physics Represents Causality

Andreas Bartels and Daniel Wohlfarth

1 Introduction

Russell's dictum that there is no place for causality in fundamental physics (Russell 1912/1913) has been revitalized in a recent debate (see, e.g., the contributions in (Price and Corry 2007)). This debate combines a rather heterogeneous collection of approaches to causality coinciding essentially in one common focal point: their approval of Russell's thesis. But Russell's thesis is seen by these authors from rather different perspectives and is appreciated for different reasons. Our critical overview of the debate (Sect. 2) shows that there are two main ways of understanding and relying on Russell's thesis that correspond to the two main arguments Russell had delivered in his 1912/1913 paper.

Russell's first reason for denying a genuine place for causality in physics was that the *asymmetry* of the causal relation – if A causes B, then it is not the case that B causes A – has no counterpart in modern theories of physics. The laws figuring in those theories make no difference whatsoever concerning the determination of the state of a system by either its past or its future states. After having declared that it is not the "same" cause producing the "same" effect, but "same" (invariant) *relations* given by differential equations that represent the so- called "law of causality" in physics, Russell claims concerning this law:

> The law makes no difference between past and future: the future "determines" the past in exactly the same sense in which the past "determines" the future. The word "determine", here, has a purely logical significance: a certain number of variables "determine" another variable if that other variable is a function of them. (Russell 1912/1913, p. 15)

Since determination by laws of physics is always a symmetric relation, there is no hope that we can find a basis for causal asymmetries in those laws. This

A. Bartels (✉) • D. Wohlfarth
Institute for Philosophy, University of Bonn, Regina-Pacis-Weg 3, 53113 Bonn, Germany
e-mail: andreas.bartels@uni-bonn.de; nullgeodaete@aol.com

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_34,

argument is the starting point for Huw Price's (2007) approach towards causality. The basis of causal asymmetry cannot be found within the laws of physics. Nor can the alleged de facto asymmetries between initial and future conditions produce any acceptable physical alternative. Since de facto asymmetries do not represent a fundamental trait of the universe, but reflect only the particular thermodynamic conditions which dominate the behavior of macroscopic systems in the low entropy era that we observe now, they cannot provide a physical basis of causal asymmetry either. Thus the basis of causal asymmetries has to lie outside physics. According to Price, it is constituted by the structure of human agency.

The second reason that has led Russell to deny the fundamental status of causality was that the determination relation of physics allegedly does not match our folk conception of causal determination. Whereas folk causal reasoning identifies one instance of an event type as the cause of an instance of another event type, physics' laws relate particular local events with global (past or future) states covering a whole cross section of the light cone of the event. We think that this does not contradict the possibility of causal reasoning on the basis of fundamental theories of physics. But we will not, in this paper, elaborate on this issue.

Thus we argue that despite the fact that both of the premises used by Russell in his arguments against the fundamentality of causality are essentially correct, neither of the conclusions follow from those premises – (1) Causality has no basis in fundamental physics and (2) Causal reasoning proceeds according to inference models that are not (and cannot) be used in physics. In this paper, however, we shall be concerned only with the first of those conclusions.

It has to be noticed that Russell attacks only one particular way in which causality could be anchored in fundamental physics, namely the way of being represented by fundamental equations. This sort of anchoring, Russell claims, is forbidden by the symmetry of determination relations as provided by fundamental equations. But there exists a further way of fundamental anchoring of causality. As we shall argue in Sect. 3, despite the time-reversal invariance of fundamental laws it is possible that the solutions of those laws are *typically*[1] *time-asymmetric*.[2] In fact, it has been proven that *almost all* spacetimes that are solutions of the field equations of General Relativity and which allow for a universal "cosmic" time parameter and, furthermore, possess a matter field are time-asymmetric.[3]

That a spacetime having a cosmic time (and therefore having spacelike hypersurfaces) is *time-symmetric* with respect to a hypersurface $t = t_S$, intuitively means that, from the hypersurface $t = t_S$, the spacetime looks the same in both temporal directions. Therefore, if a time-orientable spacetime having cosmic time is time-asymmetric, we shall not find a spacelike hypersurface $t = t_S$ which splits the

[1] In Sect. 3, we will elaborate the notion of typicality in connection with the asymmetric behavior of solutions of the field equations of GR.

[2] cf. (Castagnino and Lombardi 2009, p. 3).

[3] cf. (Castagnino et al. 2003b, p. 900 f.), (Wohlfarth 2012).

spacetime in two "halves", one the temporal image of the other one with respect to their intrinsic geometrical properties.[4] On the other hand, if we find such a hypersurface, we shall call the spacetime time-symmetric.

In Sect. 4, we will show that the result, according to which *almost all* spacetimes that are solutions of the field equations of General Relativity and which allow for a universal "cosmic" time parameter and, furthermore, possess a matter field are time-asymmetric, provides a new resource for anchoring the causal asymmetry in physics and thus diminishes the need for epistemic or even anthropocentric foundations of causality.

2 The Neo-Russellian Challenge – Overview and Critique

Huw Price sees his influential agency approach to causality as the best possible conclusion that can be drawn on the basis of our actual physical knowledge from the Russellian challenge.[5] Since Russell has convincingly shown, according to Price, that a "monarchist" conception of causality – one that bases causality on the top of the theoretical hierarchy, i.e. in the fundamental laws – is no option, whereas the other extreme "anarchistic" position would not account for the objective epistemic merits of causal reasoning, only the middle way of a "republican" option seems available. The republican holds that the notion of causality be objective, not in the sense of denoting some perspective-invariant structure of the world, but by expressing the perspectival way in which human agents conceptualize their experience of the asymmetric structure of their actions. The intervention of an agent into the course of events and the traces that result from this intervention build, from her own perspective, an asymmetric structure of its own – the asymmetric structure of causality. By what Price (2007, p. 277) calls the *fixed past principle* – the assumption that the past is typically fixed whereas the future is not – this asymmetric causal structure is aligned to the temporal asymmetry.

Since this temporal asymmetry in any region of the universe depends, according to Price, on the contingent entropy gradient in that region, the alignment of causal and temporal asymmetry does not supply causality with some fixed *universal* time sense. On the contrary, in regions in which the entropy gradient is reversed intelligent creatures would have a time-sense reversed relative to ours (Price 2007, p. 273): "... in a Gold universe our time-reversed cousins would see things differently" (Price 2007, p. 278). Therefore, the causal asymmetry cannot

[4] cf. (Castagnino and Lombardi 2009, p. 14.) In more technical terms, time-symmetry of a spacetime with respect to a spacelike hypersurface $t = t_S$ requires "time-isotropy", i.e. the existence of a diffeomorphism onto itself which reverses the temporal orientations but preserves the metric and leaves the hypersurface $t = t_S$ fixed.

[5] cf. Price (1992, 2007), (Menzies and Price 1993).

be explained by any universal time-asymmetry – rather local "arrows of time" reflect the perspectives resulting from the embedding of human agents into their particular entropic environment. In that sense, causal and temporal asymmetries are "perspectival".

We will not go into the discussion whether thermodynamics plus an initial (or final) low entropy assumption are really sufficient to explain macroscopic arrows of time in the way suggested by Price.[6] We will also not quarrel with Price's treatment of de facto temporal asymmetries, according to which over-determination of past by future conditions (he holds that this includes the asymmetry of radiation) is the result of the fact that the universe is not in a state of thermodynamic equilibrium.[7] Even if it were true that temporal experience, as expressed by the *fixed past principle*, just as de facto irreversibilities (including fork asymmetry and over-determination) exclusively depend on macroscopic, contingent thermodynamic conditions of our universe, and are thus not suited to basically explain causal asymmetry, this would not exclude the possibility that the existence of those contingent conditions itself can be traced back to some more basic (but not necessary) *time-asymmetry* of the universe. Price argues that this cannot be so, since, apart from macroscopic contingent thermodynamic conditions, the world does not entail any time-asymmetry according to theories of fundamental physics.[8] But we will refute this claim by showing that despite the symmetric structure of the fundamental equations time-asymmetry is an intrinsic trait of *almost all* general relativistic spacetimes fulfilling reasonable conditions. This provides the resource to explain causal asymmetries directly by tracing them back to the basic temporal asymmetry of the universe – thus avoiding any indirect route via thermodynamic asymmetries.

But before following this route we have to see whether there are more obvious ways to solve the problem of causal asymmetry by directly calling into question Russell's first reason to deny the fundamentality of causality: Why should it be impossible to derive causal asymmetries from laws that make no difference between the determination by past or future states?

Sheldon Smith (2000) has argued that while there are fundamental physical laws providing us with mere "functional dependencies" which fail to prefer any "direction", this does not necessarily mean that there are no laws that have the potential to represent asymmetric causal relations. Rather "causal claims emerge from the sets of functional dependencies" (Smith 2000, p. 275). This happens when such "sets of functional dependencies" are combined with dynamical equations so as to yield "a concrete differential equation of evolution type" (Smith 2000, p. 278) which has the potential to "track a causal process" (Smith 2000, p. 279). Unfortunately, Russell explicitly refers to differential equations in arguing for the

[6]For a detailed critique of a similar approach by David Albert (2000) see (Frisch 2005, 2007).

[7]cf. (Price 1992).

[8]Price uses the Gold universe as a counterexample against any intrinsic universally valid time-asymmetry.

non-causal nature of laws,[9] so Smith's answer is in any case not sufficient. On the other hand, it is rather clear what Smith has in mind. Various kinds of unequivocally causal processes like the outward radiation of an electromagnetic wave can be described by particular solutions of fundamental equations – and those solutions are expressed by differential equations. Does this not show that fundamental laws – or derivations thereof – can represent asymmetric causal relations?

It doesn't. Those differential equations entail no causal asymmetries by themselves. As Bas van Fraassen (1993) has pointed out, causal distinctions (which of two events is the cause and which is the effect) cannot be drawn by means of the formal apparatus of physical models – those formal models lack any intrinsic direction of determination. The direction rather has to be imposed on the formal apparatus, i.e. the formal model must be replaced by a *causal* model, into which a preferred direction of causal transfer has been implemented. Thus, Smith's answer wouldn't be sufficient. But perhaps models with enclosed causal interpretation will represent causality?

This is Mathias Frisch's (2012) answer. If we take theories (and their models) to be pure formalisms, then causal notions cannot be a part of them. But if we take them to include physical interpretation, it depends on the interpretation, whether causal notions are or are not part of their content. We will argue later (Sect. 3) that indeed the fundamental equations of General Relativity (GR) can be interpreted as representing causally asymmetric physical processes. In contrast to Frisch, we think that the potential to represent causal relations is not produced by adding some representational structure to an otherwise causally innocent formal model. Instead, our proposal will be that almost all models of GR have intrinsically time-asymmetric structure that builds the basis for representing causal relations.

According to Frisch, the additional structure characterizing causal models is provided by an *asymmetric fork structure*: past common causes are preferable over possible future causes because correlations between, for example, the electromagnetic fields that we observe at spatiotemporally neighboring points on Earth can be explained by the assumption of the existence of a star as the *past common source* of these correlated fields, whereas the assumption of some future cause would require some non-localized conspiracy of highly coordinated final conditions. Indeed the common cause conditions can be fulfilled by the past, not by the alternative future cause, since only the first, but not the latter is compatible with the so called *initial randomness assumption*. Initial background conditions going into our causal inferences and explanations have to be "random". Since the condition is fulfilled in only one direction – the direction we call the past-to-future direction – this principle selects one preferred causal direction.

We think that Frisch's answer concedes too much to the Neo-Russellians. It concedes that causal asymmetry is provided only by means of importing into the dynamical models genuine causal structure in form of epistemic rules governing how to construct causes out of "functional dependencies". But this is exactly what

[9] cf. (Russell 1912/1913, p. 14).

Neo-Russellians claim: Genuine causal concepts do not refer to something in nature, but are constructed by us in order to satisfy practical epistemic needs. Thus causality can only be part of physics (or of science in general) when and insofar as we embed its dynamical models into a causal inferential schema. It is exactly this message that we argue against in this paper.

3 The Time-Asymmetry of General Relativistic Spacetimes

Our aim, in the two following sections, is to show that there is a viable option for anchoring the causal asymmetry in fundamental physics: In this section, we argue that *almost all* spacetime models of general relativity are time-asymmetric (cf. Castagnino et al. 2003b, Wohlfarth 2012). This yields a new resource for anchoring the causal asymmetry in physics, and thus diminishes the need for epistemic or even anthropocentric foundations of causality. Thereby we agree with the general proposal in Maudlin's "On the Passing of Time" (Maudlin 2007a), according to which "[T]he passage of time is an intrinsic asymmetry in the temporal structure of the world" (Maudlin 2007a, p. 108), and also to Earman's *Time Direction Heresy* entailing the claim that "a temporal orientation is an intrinsic feature of space-time which does not need to be and cannot be reduced to nontemporal features" (Earman 1974, p. 20). In particular, like Maudlin we deny any attempt to reduce the direction of time to the increase of entropy.[10] But we differ with respect to how the "intrinsic asymmetry in the temporal structure of the world" is established and what it is.

Maudlin argues that the entropic atypicality[11] of microstates with respect to their backward temporal evolution (temporal evolution in backward time leads to lower entropy), in contrast to the typicality of their behavior in forward time direction, can only be explained by means of their causal production: "The atypical final state is accounted for as the product of an evolution from a generically characterized initial state." (Maudlin 2007a, p. 133) But since "[T]his sort of explanation requires that there be a fact about which states produce which" (Maudlin 2007a, p. 134) there has to be an intrinsic direction of time providing the needed causal asymmetry. Thus, in order to establish the intrinsic time-asymmetry of the universe, Maudlin takes recourse to an assumed global entropic asymmetry. But, as shown by Castagnino et al. (2003b), global entropy can only be defined in universes with a cosmic

[10] A reason against this reductionist move is provided by Maudlin: If we *define* the direction of time as the direction of entropy increase, then the Second Law has no longer empirical content: Entropy *has* to increase in forward time direction in virtue of the definition. Another objection is raised by Castagnino et al.: The entropy of the universe can only be defined under some physical conditions referring to space-time as a whole, in particular the condition that the space-time allows for a foliation into space-like hyper-surfaces, e.g. there exists a cosmic time (cf. Castagnino et al. 2003b, p. 896).

[11] See for the notion of typicality: (Maudlin 2007b) and (Goldstein 2012). Our own proposal, provided in this Section, will also make reference to this notion.

time. This seems to indicate that in establishing an intrinsic time-asymmetry of the universe, cosmological accounts have priority. Thus, we choose a cosmological account in order to establish intrinsic time-asymmetry.

We differ from Maudlin's account also concerning what this intrinsic asymmetry is. As far as we can see, there is no definite answer to this question in Maudlin's work. He claims that "all that seems to be required ... is an orientation." (Maudlin 2007a, p. 135) But it remains unclear, at least to us, what intrinsic structures of spacetime, according to his account, actually yield such orientation.

In order to avoid misunderstanding, we want to make it clear in advance that we do not aim at a *reduction* of the concept of causation to physics, i.e. we do not propose that causation can in general be *defined* in terms of fundamental physical relations as proposed by transfer theories of causation. We admit that, for instance Salmon's theory "says nothing about the asymmetry of causation."[12] We would also not propose that notorious problems like the problem of absences and non-occurrences as causes can be sufficiently tackled by means of a transfer theory of causation. Counterfactual models of causal reasoning may well play their role in physics, but we see no way to generally *define* the notions employed in counterfactual models of causation by recourse to elementary physical relations (such as energy-momentum transfer). We are exclusively concerned with causal asymmetry as a *sine qua non* condition for the existence of causal relations – however those relations may further be conceptualized in order to answer causal questions occurring in particular contexts.

Now, the first step, taken in this Section, shall be showing that time-asymmetry with respect to cosmic time is a typical property of a relevant sub-class of solutions of Einstein's field equations of GR. We take the notion of typicality to characterize a behavior of members of a set that can be shown "to hold with very rare exception".[13] In our case of interest, *most* members of a relevant class of solutions of GR field equations manifest asymmetric behavior, i.e. the asymmetric members of the class build a sub-class of *measure 1* with respect to the Lebesque measure on the solution space with respect to natural co-ordinates provided by the scaling factor, the scalar matter field and their respective derivatives.

---

[12]cf. (Hausman 1998, p. 14). Dowe (1992) when considering the problem of causal asymmetry refers to structures that are *not* intrinsic to spacetime, as entropy increase and K-meson decay. Furthermore, causal directions are not fixed by the direction of energy flow vectors. It is only a matter of convention to choose the positive flow vector as representing forward time directed physical connection between events A and B. The same physical facts are compatible with a representation in which the negative flow vector points in the direction from B to A. Thus, energy flow vectors do not, by themselves, yield causal asymmetry.

[13]cf. (Goldstein 2012, p. 60). Typical behavior is manifested, for example, by the motion of atoms of two metal rods at different temperatures that are brought into thermal contact: the motions evolve so that the temperatures in the rods equalize cf. (Maudlin 2007b, p. 287). The notion seems firstly to appear in Boltzmann's writings. In Boltzmann (1877) he argued that the evolution of a gas in line with the Boltzmann equation applies to the overwhelming majority of phase points (according to any given distribution function) – they evolve into an equilibrium state, that is, the behavior of a gas, while not inevitable, is typical (see Goldstein 2012, p. 63).

The following proof which is based on the work of Castagnino et al. ((Castagnino et al. 2003a, p. 374f), (Castagnino et al. 2003b, p. 990f.), see also Wohlfarth 2012) depends on the two following premises:

(i) The concept of cosmic time is not ruled out in the first place via the structure of the considered spacetime (as e.g. via being a non-orientable spacetime).[14]

Note that condition (i) does not yet imply the considered spacetimes to be time-asymmetric. A time symmetric universe, in which cosmic time behaves symmetrically about a symmetry point, axe or hyper array, remains still possible (see, for example, Price 1996). Hence, condition (i) does not imply the conclusion of the argument. Instead, it works like a filter, restricting the set of considered spacetimes to those for which the concept of cosmic time is applicable and not ruled out in the first place. One could also object that (i) would presuppose perfect homogeneity of the universe, while the actual universe is rather inhomogeneous. But the idealization of homogeneity just applies to uniformly co-moving galaxies with respect to which a common (cosmic) time can be defined. Non-uniform motion within the co-moving galaxies thus does not contradict the homogeneity assumption. Given the aim of this paper to show how the asymmetry of causation is based on GR, it seems perfectly adequate to consider the set of spacetimes in which cosmic time can be defined as the relevant reference-set.

The second crucial condition is the following:

(ii) The set of necessary dynamic variables contains further variables apart from the scaling factor.[15]

This condition seems plausible for physical reasons. Even if spacetimes are mathematically possible in which the only dynamic variable is the scaling factor, such spacetimes are unphysical. The set of models (the spacetimes described by a particular solution of Einstein's equation) considered here should be the set of *physically plausible* spacetimes, i.e. the set of spacetimes which contain matter and energy as dynamical entities. In spacetimes in which the only dynamical variable is the scaling factor, there is no way of representing the dynamics of the matter and energy content of the universe. Hence the set of spacetimes considered here should not include such unphysical universes.

Given conditions (i) and (ii), we will show that time-asymmetry with respect to cosmic time is a typical property of spacetimes.

---

[14]Since the condition of time-orientability guarantees that there is a consistent local time orientation for all points of spacetime, this weaker condition would be sufficient to reduce the class of spacetimes to those that have a unique local time order. But only the stronger condition of the existence of a cosmic time provides a global time function the value of which increases (decreases) along every timelike world line of the universe. Only then we can speak of "two directions of time" for the whole universe.

[15]Notice, that Price's use of the Gold universe as a counterexample to any intrinsic time-asymmetry of the universe relies on his considering the scaling factor as the only parameter characterizing the universe.

To start with: most of all *open* spacetimes are time-asymmetric. This follows from the fact that we can define the time-asymmetry of open spacetimes according to the asymmetric behaviour of the scaling factor as a function a (t) of the cosmic time t: There exists, for open spacetimes, no hypersurface $t = t_S$ such that for all t: $a(t_S + t) = a(t_S - t)$. Hence, given such open time-orientable spacetime, it is obvious that spacetime geometry looks different in both temporal directions.[16] But, GR also allows *closed* spacetimes. Hence, in order to argue that even in the set of closed time-orientable spacetime we find cosmic time-asymmetry, we have to analyse closed spacetimes in more detail (cf. Castagnino et al. 2003b, p. 900f.).

Our analysis will show that the set of time symmetric spacetimes, even in closed topologies, is a set of measure zero. More precisely, this set of time symmetric and closed solutions of Einstein's equation will turn out to have a lower dimension than the set of all closed solutions.[17]

For simplicity, consider an idealized case where the dynamics of spacetime is described by the scaling factor $a(t)$ and the scalar matter field $\phi(t)$ which depend on cosmic time t. In Hamiltonian mechanics, dynamic equations (the Hamiltonian) depend on dynamic variables and their first derivatives of the cosmic time parameter t. Thus, in this example, we have four variables in the Hamiltonian: $a(t)$, da/dt, $\phi(t)$, $d\phi/dt$. Now, analytic mechanics always allows describing one of these variables as a function of the others. Thus, for simplicity, we have chosen $a(t) = f(da/dt, \phi(t), d\phi/dt)$, where da/dt, $\phi(t)$, $d\phi/dt$ are now independent dynamic variables.

If we try to construct a symmetric spacetime, all dynamic variables must together behave in a time-symmetric manner. According to the singularity theorems in classical cosmology (Hawking and Penrose 1970; Hawking and Ellis 1973), we know that $a(t)$ has just one maximum. Next, we can choose the mathematical origin of cosmic time. For simplicity and without loss of generality, we stipulate the origin so as that $a(0)$ is the maximum value of the scaling factor. Thus, as a function of cosmic time, $a(t)$ is symmetric in relation to the axis a at the point $t = 0$, i.e. $a|_t = a|_{-t}$. Therefore, it is obvious that da/dt is symmetric in relation to the point $(t = 0; da/dt = 0)$, i.e. $\mathrm{da/dt}|_t = -\mathrm{da/dt}|_{-t}$.

[16]Note that there is one exception in classical cosmology: a static universe that also has an open topology but is symmetric (more precisely constant) with respect to the scaling factor. However, we will not consider the static solution of Einstein's equations because it requires fine tuning of the cosmological constant and the energy and matter content (and distribution) of the universe. Thus, according to classical cosmology, this solution is a special type that belongs to a subspace of solutions to Einstein's equation that has a lower dimension than the entire solution space.

[17]Spacetimes that have an open but time-symmetric (and not static) topology are open with respect to both past *and* future. We will not consider them because they require a change in the value of the cosmological constant. But, in the context of classical GR, the cosmological constant is *constant* in cosmic time. This may not be the case for full blown quantum or string cosmology, but these yet quite speculative accounts are beyond the scope of this study. In classical GR a contracting spacetime always includes a Big Crunch (see Hawking and Ellis 1973, Penrose 1979, Hawking and Penrose 1970). Thus, a spacetime cannot be open with respect to *two* directions of cosmic time if the spacetime is not static.

However, for a time-symmetric spacetime, the behaviour of $\phi(t)$ and $d\phi/dt$ *together* with da/dt must also be symmetric. Thus, in this example, we have only two possibilities for the behaviour of $\phi(t)$ and $d\phi/dt$ at the cosmic time point $t = 0$, which makes the entire spacetime time-symmetric. These possibilities are given by the triplets $\{da/dt|_{t=0} = 0,\ \phi(t = 0),\ d\phi/dt|_{t=0} = 0\}$, which is a symmetric solution of $\phi(t)$ about the $\phi-$axis at the point $t = 0$, and $\{da/dt|_{t=0} = 0,\ \phi(t = 0) = 0,\ d\phi/dt|_{t=0}\}$ being a point-symmetric solution of $\varphi(t)$ with respect to the point $(t = 0;\ \phi(t = 0) = 0)$; $\varphi|_t = -\varphi|_{-t}$.

With respect to the general definition of time-symmetry of spacetimes (Sect. 1), the t = 0-axis represents the spatial hypersurface that splits the whole spacetime into "two halves that are temporal mirror images of each other". It does not matter, for that purpose, whether the respective function is symmetric about the a-axis (hypersurface) at the point t = 0 or whether it is point-symmetric with respect to the point t = 0. In both cases, the respective function develops in the same way for an observer starting at t = 0 and going in the direction of positive values of t as for an observer starting at t = 0 and going in the direction of negative values of t. The spacetime looks physically the same in both of these temporal directions, i.e. it is a symmetric spacetime.

All symmetric solutions can be constructed using the triplets given above. Thus, we can construct a subspace of time-symmetric solutions: *span* $\{(da/dt|_{t=0} = 0,\ \phi(t = 0),\ d\phi/dt|_{t=0} = 0),\ (da/dt|_{t=0} = 0,\ \phi(t = 0) = 0,\ d\phi/dt|_{t=0})\}$. The complete space for solutions of the dynamic equation, however, is given by *span* $\{(da/dt,\ 0,\ 0),\ (0,\ \phi,\ 0),\ (0,\ 0,\ d\phi/dt)\}$

Thus, time-symmetric behaviour of a spacetime occurs only in a subset of solutions having a lower dimension than the entire set of solution *even* if we consider closed spacetimes.

The argument given above has shown that, assuming that cosmic time is definable and that we have more than the scaling factor to describe the dynamics of spacetime, time-asymmetry in terms of cosmic time is a typical property of the most idealized model. But this also holds true if we leave the very simple toy model of the example and add more dynamical variables, because the calculation in those cases would proceed in the same manner. The entire space of the solutions always has a higher dimension than that of the subset of time-symmetric solutions.

Despite the fact that the fundamental Einstein equation is time reversal invariant, this does not imply that the models of General Relativity are time-symmetric. Instead, we have shown that *almost all* models, in which cosmic time can be defined, are time-asymmetric. Time-asymmetry is a typical property for the relevant set of solutions of the GR field equations.

Every time-asymmetric model of GR has a time-reversed model which is also a solution of the Einstein equation. Thus one could object that causal asymmetry would not follow from any consideration which uses the time-asymmetry of models of GR.

But this objection fails.[18] The two models *f(t)* and *f(−t)* can be shown to represent the same physical world. We argue for that in the following three steps:

(a) Since the models are isomorphic, $f(t)$ does not possess intrinsic properties that are not possessed by $f(-t)$.
(b) The models describe spacetime as a whole. This implies that there is no time parameter (or any other physical parameter) outside the range of the geometrical objects $f(t)$ and $f(-t)$. The models are thus not related to a physical environment.
(c) The combination of conditions (a) and (b) shows that two models $f(t)$ and $f(-t)$ do not differ in any intrinsic *or* extrinsic property. Thus, the models represent the same physical world.

It should be noticed that it is not only the case that *we* just cannot *distinguish* between the two solutions. The argument shows that there cannot be any property distinguishing between the two models.

Nevertheless, as of yet we have only shown that almost all models of General Relativity are globally time-asymmetric (with respect to cosmic time). Since there is no obvious way to connect time-asymmetry of cosmic time with the time behaviour of physical processes, we have not yet shown that those models have the potential to represent asymmetric causal processes.

Hence, the second step in our argumentation will be to show that such a specific kind of global time-asymmetry has crucial consequences for the proper time parameter of all world lines in such a spacetime. More precisely: We shall show that, at least in spacetimes similar to ours, the proper time parameter for individual world lines behaves asymmetrically. This local time-asymmetry, as we shall argue, provides the basis for the representation of causal relations according to GR.

4 The Way to Causal Asymmetry

Our first aim, in this Section, is to deduce a temporal asymmetry of proper times from the global time-asymmetry[19] and to analyse the set of additional assumptions needed to proceed in that way.

Regarding this issue, it is well known (cf. Earman 1974) that we can use a non-vanishing, continuous timelike vector field on a time-orientable spacetime to distinguish between the semi-light cones:

[18] We follow here the argumentation of (Castagnino and Lombardi 2009, p. 18).

[19] We follow here the mathematical procedure of Castagnino and Lombardi (2003a, p. 376f; 2009, p. 19 f.). But we will not agree with the view of Castagnino et al., according to which positive local energy flow as constructed in this procedure selects a substantial future direction of time and thus defines a local arrow of time.

> Assuming that spacetime is temporally orientable, continuous timelike transport takes precedence over any method (based on entropy or the like) of fixing time direction; that is, if the time senses fixed by a given method in two regions of spacetime (on whatever interpretation of regions you like) disagree when compared by a means of transport that is continuous and keeps timelike vectors timelike, then if one sense is right, the other is wrong. (Earman 1974, p. 22)

In the following, we provide the conditions for the construction of a non-vanishing continuous timelike vector field which, since it describes energy-momentum flow, is a plausible candidate for representing causal connections in spacetime. This mathematical object turns out to have exactly the physical meaning that we desire in order to anchor the concept of causation within fundamental physics. Furthermore, it can be used, according to the method mentioned above, to fix a local time sense. But one should not assume, at this point, that this already gives us the distinction between a local "past" and "future". All that this object provides us with is a method to describe local causal connections that follow a distinct local time sense which can consistently be extended over the whole spacetime – provided that a local time sense had been selected at one point. But how can the two possible local time senses be physically distinguished?

It is exactly at this point that *global* time-asymmetry comes into play. Given the physical difference between the two cosmic time-directions in almost all spacetimes (cf. Sect. 3) the distinction between the two possible local time senses at a certain point of spacetime also gets a substantial physical meaning: One semi light-cone at the point contains all the timelike vectors pointing in one of two geometrically different cosmic time directions, whereas the other semi light-cone contains all the timelike vectors pointing in the other cosmic time-direction. Thus the global time-asymmetry is transferred to the local realm, with the effect that the local time senses become distinct with respect to the physically distinct global time directions. But notice that the physical distinctiveness still does not give us an answer to the "which-is-which"-question: We still cannot tell *which of the two local time senses is the "past" and which is the "future" time sense*. What we have achieved instead is an answer to the question of how the asymmetry of local causal relations is anchored in global time asymmetry.

The first step in order to construct the desired non-vanishing timelike vector field is now to identify possible physical candidates. As we will see below, for this task it turns out to be useful to consider the energy–momentum tensor[20]:

$$T_{\mu\nu} = 1/8\pi \left(R_{\mu\nu} - 1/2 g_{\mu\nu} R - \Lambda g_{\mu\nu} \right) \tag{1}$$

Since the components of $T_{\mu\nu}$, as they occur in Eq. (1), do not play the role of a continuous, non-vanishing timelike vector field in general, we have to add two conditions for $T_{\mu\nu}$, namely

[20]Here $R_{\mu\nu}$ is the Ricci tensor, R the Ricci curvature, Λ is the cosmological constant and $g_{\mu\nu}$ the metrical tensor.

(I) $T_{\mu\nu}$ is a type one energy–momentum tensor[21]
(II) $T_{\mu\nu}$ satisfies the dominant energy condition $T^{00} \geq |T^{\mu\nu}|$ for any orthonormal basis

If condition I is satisfied, then we can write Eq. (1) in the form

$$T_{\mu\nu} = s_0 V_\mu{}^0 V_\nu{}^0 + \sum_{i=1}^{3} s_i V_\mu{}^i V_\nu{}^i \tag{2}$$

$\{V_\mu{}^0, V_\nu{}^i\}$ being an orthonormal tetrad and, as in the standard notation, $V_\mu{}^0$ being a timelike and $V_\nu{}^i$ a spacelike vector with $i \in \{1,2,3\}$.

If condition (II) is satisfied as well, then it follows that $s_0 \geq 0$. Therefore, if s_0 is not zero:

(A) Conditions (I) and (II) are fulfilled in almost all world models considered in classical cosmology.
(B) $V_\mu{}^0(x)$ (where x represents the spacetime coordinates) is a continuous, non-vanishing timelike vector field. Moreover, $T^{0\mu}$ can be interpreted as the physical energy flow, described by a continuous, non-vanishing timelike vector field.[22]

According to B, in all time-asymmetric spacetimes satisfying the conditions I and II, we find a physical vector field on which the time-*asymmetric causal connection* between events can be based. We will say that events C and E are *causally connected* iff there is a time-asymmetric energy flow from C to E.

It might be objected that the foregoing explication of causal connections implies that the causal asymmetry is to be *defined* by time-asymmetry. Indeed, according to our view the asymmetry of the causal connection is derived from global time-asymmetry that is transferred itself to the local realm. But in contrast to some merely conventional stipulation, this derivation provides the causal asymmetry with specific physical significance: the direction into which the causal "arrow" points (from C to E) differs substantially from the opposite direction by being aligned to some distinguishable cosmological time direction in which the geometry "looks" different as compared with the opposite direction. Thus, the *causal directions are physically distinct from each other.*[23]

[21]This means that the tensor can be described in normal orthogonal coordinates. See Hawking and Ellis 1973 and also Eq. 2 for the mathematical meaning of "type I" or "normal" in this context.

[22]This interpretation appears to be canonical in the context of general relativity, but there are exceptions that show that this understanding of $T^{0\mu}$ is not valid in general. The exceptions come into play by considering quantum effects. Critical points are, for example, the Casimir effect or squeezed vacuum or Hawking-evaporation. (see e.g. Visser 1996; Barceló and Visser 2002)

[23]We thereby agree with the view of Faye (1996, 2002) that the causal direction from cause to effect is *objective*: it can be distinguished from the opposite direction by physical facts. But unlike Faye we argue that these facts are such as to establish, in the first instance, a global *temporal* asymmetry. The *causal* asymmetry is derived from this global temporal asymmetry.

What we have established by now may be called a *weak* causal arrow: Causal relations between events have a substantial (not conventional) time-direction that is in line with one of the global time-directions which are substantially (not conventionally) different in virtue of their particular geometrical characteristics. But we get no definite answer to the question *which* global direction is selected as the "past" (or "future") direction, with "past" and "future" having their common meaning as manifested in daily experience ("fixed past principle"). To get an answer to that question would require being able to solve the problem of a *strong* causal arrow – being able to single out a unique global arrow of time which could then be transferred to the local realm with the effect that the selected directions coincide with the experienced past and future. Instead of this, what we actually have achieved is only a *distinction* between two global time directions without any method to tell which is which. We have thus provided only a solution to the problem of the weak causal arrow – to the problem "of finding a substantial asymmetry of time that allows us to distinguish between both temporal directions." (Castagnino and Lombardi 2009, p. 5)

Unlike the view we propose in this paper Castagnino and Lombardi (2009) pursue the more ambitious project to establish a strong arrow: "... we can meaningfully say that future is the temporal direction of the positive local energy flow: the local flow of energy emitted at x points to the future ..." (Castagnino and Lombardi 2009, p. 21). But a pure stipulation, according to which the energy flow vector is contained in the conventionally labeled "future" light cone would be insufficient for achieving any solution to the strong arrow problem. In order to yield such a solution, it would be necessary to establish a *physically substantial* future direction. This, in turn, requires an account of its own, according to which the global time directions cannot just *differ* by their peculiar geometrical evolution, but one direction can be *singled out* to be the global direction of energy flow in the universe, and therefore the unique global future. This unique global time direction would then define substantial local future directions, according to which energy flow spreads out into space. Retarded wave solutions are thus selected by being those solutions that describe such a spread out into a substantial local future. Castagnino et al. have indeed proposed such an account (cf. Castagnino et al. 2003a, p. 380f.).

Since the decisive assumption in that account is that energy flow (and therefore causal processes) in the universe define some universal arrow of time, we think that the whole argument collapses into a *petitio*. The real work that would have to be done, in order to yield a strong arrow would be such: It would have to be shown that exactly *one* of the distinct global time-directions, because of the peculiar nature of the geometrical evolution in that direction, is the direction into which retarded electromagnetic waves spread out. The energy flow connected to them would then define a local future. As a consequence, local future and causal arrow would indeed depend on the nature of geometry. But that programme has not even been started. The advantage of our account of causal asymmetry may be that it shows how far one can go along with actual physics and conceptual analysis. Thereby the precise frontiers of future work come into sight.

5 Conclusions

We have shown that most of all spacetimes of GR are globally time-asymmetric. This global time-asymmetry is transferred to the local light-cone structure. Thus a continuous timelike vector field which has the physical meaning of energy flow and can thus be considered to represent causal connections between events, defines a local time sense with respect to physically distinct global time directions. We suggest that this is the physical basis of the asymmetry of the causal relation. Thus, we have defeated the Neo-Russellian claim, according to which the asymmetry of causation cannot be derived from physics. Causation, contrary to those claims, has a prominent place in physics, at least with respect to fundamental causal connections.

References

Albert, D. 2000. *Time and chance*. Cambridge, MA: Harvard University Press.

Barceló, C., and M. Visser. 2002. Twilight of the energy conditions? Preprint gr-qc/0205066.

Boltzmann, L. 1877. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respective den Sätzen über das Wärmegleichgewicht. *Sitzungsberichte der Akademie der Wissenschaften zu Wien* 76: 373–435.

Castagnino, M., and O. Lombardi. 2009. The global non-entropic arrow of time: From global geometrical asymmetry to local energy flow. *Synthese* 169: 1–25.

Castagnino, M., L. Lara, and O. Lombardi. 2003a. The cosmological origin of time asymmetries. *Classical and Quantum Gravity* 20: 369–391.

Castagnino, M., O. Lombardi, and L. Lara. 2003b. The global arrow of time as a geometrical property of the universe. *Foundations of Physics* 33(6): 877–912.

Dowe, P. 1992. Process causality and asymmetry. *Erkenntnis* 37: 179–196.

Earman, J. 1974. An attempt to add a little direction to 'the problem of the direction of time'. *Philosophy of Science* 41: 15–47.

Faye, J. 1996. Causation, reversibility and the direction of time. In *Perspectives on time*, ed. J. Faye, U. Scheffler, and M. Urchs, 237–266. Dordrecht: Kluwer.

Faye, J. 2002. When time gets off track. In *Time, reality & experience*, ed. C. Callender, 1–17. Cambridge: Cambridge University Press.

Frisch, M. 2005. Counterfactuals and the past hypothesis. *Philosophy of Science* 72: 739–750.

Frisch, M. 2007. Causation, counterfactuals and entropy. In *Causality, physics, and the constitution of reality: Russel's republic revisited*, ed. H. Price and R. Corry, 351–396. Oxford: Oxford University Press.

Frisch, M. 2012. No place for causes? Causal skepticism in physics. *European Journal for the Philosophy of Science* 2(3): 313–336.

Goldstein, S. 2012. Typicality and notions of probability in physics. In *Probability in physics*, ed. Y. Ben-Menahem and M. Hemmo, 59–71. Berlin: Springer.

Hausman, D. 1998. *Causal asymmetries*. Cambridge: Cambridge University Press.

Hawking, S., and G. Ellis. 1973. *The large scale structure of space-time*. Cambridge: Cambridge University Press.

Hawking, S., and R. Penrose. 1970. The singularities of gravitational collapse and cosmology. *Proceedings of the Royal Society A* 314: 529–548.

Maudlin, T. 2007a. *The metaphysics within physics*. Oxford: Oxford University Press.

Maudlin, T. 2007b. What could be objective about probabilities? *Studies in History and Philosophy of Modern Physics* 38: 275–291.

Menzies, P., and H. Price. 1993. Causation as a secondary quality. *The British Journal for the Philosophy of Science* 44: 187–203.

Penrose, R. 1979. Singularities and time asymmetry. In *General relativity: An Einstein centenary survey*, ed. S.W. Hawking and W. Israel, 617–629. Cambridge: Cambridge University Press.

Price, H. 1992. Agency and causal asymmetry. *Mind* 101(403): 501–520.

Price, H. 1996. *Time's arrow and archimedes' point*. Oxford: Oxford University Press.

Price, H. 2007. Causal perspectivalism. In *Causality, physics, and the constitution of reality: Russell's republic revisited*, ed. H. Price and R. Corry, 250–292. Oxford: Oxford University Press.

Price, H., and R. Corry (eds.). 2007. *Causality, physics, and the constitution of reality: Russell's republic revisited*. Oxford: Oxford University Press.

Russell, B. 1912/1913. On the notion of cause. *Proceedings of the Aristotelian Society* 13: 1–26.

Smith, S. 2000. Resolving Russell's anti-realism about causation: The connection between causation and the functional dependencies of mathematical physics. *The Monist* 83(2):274–295.

Van Fraassen, B. 1993. Armstrong, Cartwright, and Earman on laws and symmetry. *Philosophy and Phenomenological Research* 53(2): 423–429.

Visser, M. 1996. *Lorentzian wormholes: From Einstein to Hawking*. New York: Springer.

Wohlfarth, D. 2012. A new view of 'fundamentality' for time asymmetries in modern physics. Proceedings of the EPSA 11 conference in Athens, October 2011: Recent progress in philosophy of science: Perspectives and foundational problems. New York: Springer.

Anchoring Causal Connections in Physical Concepts

Mario Hubert and Roland Poellinger

1 Grounding (Not Reducing) Causality

The question about causality seems to be ubiquitous in the natural sciences and especially in physics. Why and how did the universe evolve? Why is there life on earth? What caused a particle to have that particular trajectory? But if we look closer and examine our physical theories in detail, the issue of how exactly to answer such questions is not that clear. Bertrand Russell was famous for stressing that our fundamental physical theories do not tell us anything about causal relations in their domain since the equations figuring in these theories are time-reversal invariant, and causation seems to be parasitic on time-asymmetry.[1] Even present-day philosophers of science, such as John Norton, do not grant causality any role in physics, because there is no *universal principle of causality that holds true of our science* unless physical theories are *restricted to appropriately hospitable domains*.[2]

[1] Cf. Russell (1912).

[2] Cf. Norton (2007).

M. Hubert (✉)
Section de philosophie, Faculté des lettres, Université de Lausanne, 1015 Lausanne, Switzerland
e-mail: Mario.Hubert@unil.ch

R. Poellinger
Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich, Ludwigstr. 31, 80539 Munich, Germany
e-mail: R.Poellinger@lmu.de

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_35,

In contrast, physicists themselves normally adhere to an intuition very close to what is known as causal fundamentalism, which claims that the job of all of physics is to uncover the prevalent causal relations in nature. In one of the most famous modern textbooks on classical electrodynamics one finds the following pointed affirmation –

> [. . .] the most sacred tenet in all of physics: the principle of causality.[3]

Neo-Russellians, such as Huw Price, in general still deny causal relations any role in physics, but acknowledge their importance in the special sciences and in common sense reasoning. However, there are also philosophers like Mathias Frisch, who take up a position between the Neo-Russellians and the causal fundamentalists. Frisch for example claims that causal notions can play a role in physics and do play a role in certain domains when one adds an interpretation to the formal apparatus of a physical theory, which finally allows us to apply the machinery of causal reasoning again.

A seemingly imperative task prior to the quest for causal relations in physics is to try to establish the direction of time within the physical theories under consideration, e.g., with the help of entropy and the second law of thermodynamics (thermodynamic arrow of time), the expansion of the universe (cosmological arrow of time), or the direction of retarded waves (radiative arrow of time). Although these explanations of the direction of time are not unproblematic, the question we are concerned with here is this one: how can we use the time-asymmetry within physical theories in searching for causality?

Following this route, Andreas Bartels and Daniel Wohlfarth set out to show in their paper *How Fundamental Physics Represents Causality*[4] that there is place for causality in one of our most successful fundamental theories, namely General Relativity (GR). As time-asymmetry is supposed to be a necessary condition for causal relations, their strategy is first to show in what sense General Relativity is time-asymmetric. Having done that, they provide a description on how to build causality on it. Throughout the paper they emphasize that the time-reversal invariance of the equations does not imply that the solutions must be time-symmetric.

This point is often ignored; so we explicitly want to give the definitions. An *equation* – respectively a law if the law is mathematically formulated as an equation – is said to be time-symmetric (time-reversal invariant) iff for any solution $f(t)$ of that equation $f(-t)$ also presents a solution. We call a *function* $f(t)$ – which is in the most interesting cases a solution of a law-like equation – time-symmetric iff there is a t_0 such that $f(t_0 + t) = f(t_0 - t)$ for all t.

The construction of time-asymmetry by Bartels and Wohlfarth relies on the work of Mario Castagnino,[5] who considers models of space-time which can be described by the scale factor and a matter field. It turns out that these models are typically time-asymmetric w.r.t. cosmic time.

---

[3] Cf. Griffiths (1999, p. 425).

[4] Cf. Bartels and Wohlfarth (2014).

[5] Cf. Castagnino et al. (2003a,b,c) and Castagnino and Lombardi (2009).

The next task is to deduce *local* time-asymmetry by constructing a non-vanishing, continuous, time-like vector field on the time-orientable space-time. Starting from the energy-momentum tensor

$$\mathrm{T}_{\mu\nu} = \frac{1}{8\pi}\left(\mathrm{R}_{\mu\nu} - \frac{1}{2}\mathrm{g}_{\mu\nu}\mathrm{R} - \Lambda\mathrm{g}_{\mu\nu}\right), \tag{1}$$

with $\mathrm{R}_{\mu\nu}$ the Ricci tensor, R the Ricci curvature, Λ the cosmological constant, and $\mathrm{g}_{\mu\nu}$ the metrical tensor, we get (after imposing further restrictions on the energy-momentum tensor)

$$\mathrm{T}_{\mu\nu} = s_0 V^0_\mu V^0_\nu + \sum_{i=1}^{3} s_i V^i_\mu V^i_\nu, \tag{2}$$

such that $s_0 \geq 0$, $s_i \in \mathbb{R}$, V^i_ν are space-like vector fields for all $i \in \{1, 2, 3\}$, and V^0_μ is a time-like vector field on the space-time manifold. Making explicit the space-time coordinate x, $V^0_\mu(x)$ is the supposed non-vanishing, continuous, time-like vector field that establishes the necessary local time-asymmetry by being a means of distinguishing between the semi-light-cones on every space-time point. Furthermore, there is a physical quantity connected to $V^0_\mu(x)$: T^0_μ is interpreted as the energy flow in the direction of $V^0_\mu(x)$.

Finally, causality may enter the arena as stated in the paper:

> [W]e find a physical vector field on which the *time-asymmetric causal connection* between events can be based. We will say that events C and E are *causally connected* iff there is a time-asymmetric energy flow from C to E.

This is the crux of how Bartels and Wohlfarth anchor causality in General Relativity. We will discuss their proposal in the following. In Sect. 2 we introduce the most prominent theories of causation and ask what concept of causation Bartels and Wohlfarth use. We then examine in Sect. 3 the details of how they ground causality in General Relativity upon formalizing the above quotation. Section 4 presents a short summary and outlook.

2 Concepts of Causation

Outside of physics (many times on the border), philosophers have been thinking systematically about cause and effect since the very beginnings and even more rigorously with the introduction of mathematical methods and formal semantics into the discipline in the last century. Epistemology and philosophy of science at once had the means to shape prevailing problems in symbolic form, express achievements with scientific stringency, and sort issues within formal theories from questions about intuitions and basal premisses. Select approaches to characterizing causal

relations shall be outlined in the following to give an overview over the problems a causal theorist is facing when casting plausible analysis in formal structure.

David Hume's famous quote may be seen as the point of reference for many formal theories – he makes out an essentially structural unifying feature of causal relations when he claims in 1748 that "*[w]e may define a cause to be an object followed by another, and where all objects, similar to the first, are followed by objects similar to the second*." This portion of *An Enquiry about Human Understanding* (Sect. VII) becomes the corner stone of the regular and ultimately probabilistic analysis of causation connected with the names of, e.g., Good and Reichenbach or also Suppes, who explicitly builds the direction of time into his account to express our intuitions about the *temporal* asymmetry in formal manner: a cause must necessarily be correlated with its effect *and* precede it.

Nancy Cartwright is well-known for her critique of a purely mathematical, *thin* characterization of causation, especially of the one based merely on regularities or correlations – she ultimately emphasizes the experimenter's knowledge about the experimental setup and ties methodology and interpretation together as two sides of the same causal coin. In her eyes, transferring causal knowledge from narrowly defined lab conditions to situations of larger scale or everyday experience cannot follow one single principle, on the contrary, it must naturally be as diverse (though maybe family-like) in character as the methodology applied in the first place itself. What is important for our discussion, though, is her standpoint that causality cannot be a monolithic, fundamental concept due to its multifarious nature.

The proponents of an interventionist accounts of causation agree with Cartwright on one central thing: they acknowledge the problems of a mere statistical analysis of causality. The interventionist theorist wants to do better by adding an element of counterfactual analysis to the probabilistic framework without being accused at the same time of metaphysically overloading our mostly solid notion of causation. Structural hypothetical interventions, utilized for the test of causal efficacy and formally expressed as fixing the value of a (random) variable in a Bayesian network, connect with a scientist's practice and mathematical toolbox at the same time – without causation coming under the suspicion of being reduced to an anthropomorphic concept. Causes are expressed as difference-makers in given circumstances.

Interventionist theorists as Judea Pearl clearly localize causal relationships with underlying physical mechanisms on the objective side of things and dismiss a subjective or even epistemic reading. The interventionist framework is open to this reading, nevertheless. Ultimately, deciding upon the set of variables considered illuminating for the analysis to be conducted is obviously a subjective (sometimes highly pragmatic) process that may differ from one epistemic agent to the next even if performed in compliance with rational standards. Jon Williamson, for example, is opting for a fruitful exploration of causal graphs as bearers of epistemic contents and direct enablers of meaningful communication. Cases of causation by omission, the distinction whether a result was actively obtained or passively not prevented, as well as the question how to determine accountability on the basis of causal efficacy can be made transparent in an epistemic account without much hassle.

Now, when Bartels and Wohlfarth set out to search for fundamental prints of causation in physical grounds, which of the many readings of causation do they have in mind, or in other words: what might constitute a good candidate set of features for them that makes a causal relation causal?

Just as Bartels and Wohlfarth some philosophers have tried to take a different perspective and approach the task of formalizing intuitions about causality from a point of view closer to physics. The causal powers theorist is straightforwardly asking the question why not to introduce causality as a basic power and ascribe essential causal capacities to objects of reality. *Dispositions* are meant to be necessarily separate from their token instantiations, but at the same time linked to those instantiations of themselves through a necessary causal relation: causal powers (as Popper's *propensities*) are seen as enduring states with the disposition to objectively produce events or states by singularly contributing observable quantities to their manifestations. Problems with this account arise as soon as we ask about the nature of the connection between those powers and their manifestations. And: can one really postulate a certain disposition if it, for example, never manifests itself? And if we want time-asymmetry to be built into the expression of a causal relation: is there a way to understand the directedness of powers as necessary causal directedness from cause towards effect?

Processes seem to be another promising fundamental building block candidate for a theory of causation. At the core of process theories lies the explication of world lines and their intersection, understood as more basic than the causal relation between events. Phil Dowe extends Wesley Salmon's material work[6] by introducing exchanged conserved quantities, such as linear momentum, mass-energy, or charge, to make it empirically applicable.

Dowe's theory relies on the following two propositions[7]:

1. A *causal interaction* is an intersection of world lines which involves exchange of a conserved quantity.
2. A *causal process* is a world line of an object which possesses a conserved quantity.

A *world line* is the collection of points in space-time that represents the history of an object. A *process* is understood as the world line of an object, regardless of whether or not it possesses any conserved quantities. As we are here concerned with causation in fundamental physics, an *object* is anything found in the ontology of a fundamental physical theory, e.g., particles, waves, or fields. A *conserved quantity* is any quantity that is universally conserved in our actual physical world. Our current fundamental physical theories tell us what these quantities are (most prominently energy) and by which laws they are governed. An *intersection* simply is the overlapping of two or more processes in space-time – it consists of all space-time points common to both (or all) processes. An *exchange* occurs when

[6]Cf. Salmon (1997).

[7]Cf. e.g. Dowe (2000).

at least one incoming and at least one outgoing process undergo a change in the value of the conserved quantity – in this case, "incoming" and "outgoing" are defined with respect to the light-cone structure of space-time.

With these definitions at hand one can finally state that events C and E are causally connected (connected by a causal relation) iff a continuous series of causal processes and mediating interactions can be traced between them. In this very coarse first draft of a naïve process theory, however, it seems that routine analysis returns too many causes, which can only be reduced again by utilizing extra-theoretical knowledge or assumptions (against the original goal of theoretically objectivizing the notion of a causal process). Moreover, although the conserved quantity theory seeks to ground causality in physics it seems not to be compatible with GR.[8] The most important reason for this is the lack of an energy conservation law for most models of GR. In general, energy is not a conserved quantity in this theory, and for this very reason the notion of a physical process becomes meaningless if it is to be built upon any definition of energy.

One can of course reply that the conserved quantity theory is applicable to isolated physical systems where it is possible to reclaim conservation of energy, and restrict oneself to these cases. However, isolated systems are idealizations that are an abstraction of our actual world and, strictly speaking, cannot be found therein. Current research raises further questions along these lines and beyond our considerations in this paper: is it possible that our actual world can perhaps be described by some space-time that allows for a conservation of energy within GR? We shall move on to Bartels and Wohlfarth's bridge building between causal claims and physical terms.

3 Bridging Causality and Energy Flow

In order to elicit the energy-flow T^0_μ from GR, Bartels and Wohlfarth impose a myriad of restrictions on the space-time model they use. They only consider models that can be described by the scale factor plus a matter field. Almost all of these models turn out to be time-asymmetric w.r.t. global time. To establish local time-asymmetry and consequently energy-flow they introduce some further – rather technical – constraints on the energy-momentum tensor. Is the resulting space-time model a model of the universe we live in? Is causality thus grounded in the physical world or just in a special solution of GR? Is it also possible to ground causality in space-time models other than the ones described by a scale factor?

Bartels and Wohlfarth's central statement finally is the formulation of the link between two events' causal and physical relation. Their ultimate goal is to associate the time-asymmetric causal connection between events with physical concepts in

[8]The arguments for this proposition are taken from Lam (2005) referring to Curiel (2000) and Rueger (1998).

the aforementioned claim that "events C and E are *causally connected* iff there is a time-asymmetric energy flow from C to E." This statement summarizes Bartels and Wohlfarth's view that causal asymmetry is not to be conventionally defined by concepts anchored in physics, but that it is rather "derived from global time-asymmetry" and thus intrinsically physically endowed with two distinct directions.

We read the formulation "events C and E are causally connected" as "event C causally contributes to event E" in order to avoid the interpretation of C as the true, the sole, or the actual cause of E and, at the same time, to avoid talk of prevented or potential causation and the like. To us *causally connected* is just a very weak notion conveying that the event C plays a certain (yet to be determined) causal role for the occurrence of E. However, Bartels and Wohlfarth neither explain what they mean by *causal connection* nor state what notion of causality they intend to ground in physics. Instead, they introduce the following restrictions:

> In order to avoid misunderstanding, we want to make it clear in advance that we do not aim at a *reduction* of the concept of causation to physics, i.e. we do not propose that causation can in general be *defined* in terms of fundamental physical relations as proposed by transfer theories of causation. [...] We would also not propose that notorious problems like the problem of absences and non-occurrences as causes can be sufficiently tackled by means of a transfer theory of causation. Counterfactual models of causal reasoning may well play their role in physics, but we see no way to generally define the notions employed in counterfactual models of causation by recourse to elementary physical relations (such as energy-momentum transfer).

Clearly, Bartels and Wohlfarth do not claim that all causal relations can be explained as relations described by fundamental physics, nor do they want to include omissions or counterfactual statements in their examination. They continue:

> We are exclusively concerned with causal asymmetry as a *sine qua non* condition for the existence of causal relations – however those relations may further be conceptualized in order to answer causal questions occurring in particular contexts.

Causal reasoning in physical contexts fundamentally relies on causal asymmetry: if C causes E, then E does not cause C. Coupling this causal asymmetry with temporal asymmetry we get that C is temporally prior to E if C causes E. It is this line of reasoning, which seems to be Bartels and Wohlfarth's motivation for grounding causal relations in the time-asymmetry of certain models of GR.

Now, in a first attempt to carve out the logical structure of Bartels and Wohlfarth's central claim we might look at the following formula:

$$\forall C, E \in \mathcal{E} \left(C \twoheadrightarrow E \iff \exists f (C \overset{f}{\rightsquigarrow} E) \right) \tag{3}$$

with $\mathcal{E}$ the global set of events, '$\twoheadrightarrow$' representing the time-asymmetric causal connectedness between two events, and '$\overset{f}{\rightsquigarrow}$' representing the time-asymmetric energy flow f between two events.

Now, as far as we can see all causal theories agree with '$\Longleftarrow$', in fact, most (if not all) formal theories of causation will use this direction as one of the crucial

applicability benchmarks – in general: the postulation of energy flow between two events supports the intuition that these two events are bound causally in a push-pull way, even if C is not an actual, maybe only a contributing cause of some effect E.

The other direction, though, seems to raise some questions: the set $\mathcal{E}$ will have to be restricted in a suitable manner to allow for such an inference. As said above, omissions and non-occurrences are excluded by Bartels and Wohlfarth themselves, who admittedly aim for some (limited) concept of *physical causation*.

Now, any potential restrictions should be explicitly expressible – we might consequently modify our formula above by imposing a set of conditions Γ on C and E:

$$\forall\, C, E \in \mathcal{E} \left(\Gamma \Longrightarrow \left(C \twoheadrightarrow E \iff \exists f (C \overset{f}{\rightsquigarrow} E) \right) \right) \tag{4}$$

What might Γ stand for? If the premises of the right-hand side (e.g., the existence of a distinguished physical vector along which the direction of energy flow is to be aligned) turn out to constitute the basis for the left-hand side as well, the biconditional thus formulated might become insubstantially *thin* in the end. If by narrowing down potential C–E pairs only process-like connected Cs and Es remain, we are essentially left with postulating 'measurable exchange of conserved quantities' as an explication of '$\twoheadrightarrow$', or simply put: "energy flow". The question remains: if the proposed biconditional is not a definition, what intuitions ought '$\twoheadrightarrow$' to capture above and beyond '$\exists f(\cdot \overset{f}{\rightsquigarrow} \cdot)$'?

4 Arrows and Targets

Bartels and Wohlfarth's paper provides an interesting answer to Russell's fundamental critique in considering time-asymmetric solutions of a time-symmetric physical law on which causal asymmetry might then be based in turn, "a *sine qua non* condition for the existence of causal relations", in any case a condition that usually has the status of an unquestioned, almost axiomatic precondition for causal analysis. The authors consequently aim to establish a *weak* causal arrow:

> Causal relations between events have a substantial (not conventional) time-direction that is in line with one of the global time-directions which are substantially (not conventionally) different in virtue of their particular geometrical characteristics.

In their investigations the authors make out the asymmetry of global time upon which causal relations might be based, finally. However, this should be distinguished from a much stronger claim: the problem of the *strong* causal arrow, where one has to show additionally which global direction is the future and which the past direction we experience in daily life. Using this distinction one has to be attentive not to mix problems concerning the arrow of time and causality, though they are surely associated, as we could see in this paper.

We agree with Bartels and Wohlfarth's critique of Castagnino and Lombardi (2009) at the end of their paper: it is too simple to merely stipulate that the future direction coincides with the direction of *positive* local energy flow. Instead, one should state that the geometrical development of space with respect to global time characterizes future and past. This can then be utilized to define local future and local past and to finally anchor causal connections in physical concepts – in models where the transition from global time to a local one is possible. However, although the directionality of time might well be the common thread of asymmetric energy flow and the asymmetry of causation, it still remains unclear in the paper why exactly '$\multimap$' should be named "causal", and if so, what exactly the meaning of this attribute (undefined until last) might actually be.

Looking at the big picture, it will be inevitable to determine whether our actual world can be described in accordance with the physical restrictions Bartels and Wohlfarth impose on the general relativistic space-time model they use in the first place, such that their causal notion can truly be based in nature. We find the idea of geometrically grounding causation in physics by formally building on time-asymmetry very fruitful – next targets could be the problem of the strong causal arrow as well as the clarification of how this construction connects with other theoretical approaches towards causation.

References

Bartels, A., and D. Wohlfarth. 2014. How fundamental physics represents causality. In *New directions in the philosophy of science*, ed. M.C. Galavotti, et al. New York: Springer.

Castagnino, M., and O. Lombardi. 2009. The global non-entropic arrow of time: From global geometrical asymmetry to local energy flow. *Synthese* 169(1): 1–25.

Castagnino, M., L. Lara, and O. Lombardi. 2003a. The cosmological origin of time asymmetry. *Classical and Quantum Gravity* 20(2): 369.

Castagnino, M., L. Lara, and O. Lombardi. 2003b. The direction of time: From the global arrow to the local arrow. *International Journal of Theoretical Physics* 42: 2487–2504.

Castagnino, M., O. Lombardi, and L. Lara. 2003c. The global arrow of time as a geometrical property of the universe. *Foundations of Physics* 33: 877–912.

Curiel, E. 2000. The constraints general relativity places on physicalist accounts of causality. *Theoria – Segunda Época* 15(1): 33–58.

Dowe, P. 2000. *Physical causation*. Cambridge: Cambridge University Press.

Griffiths, D. 1999. *Introduction to electrodynamics*. Upper Saddle River: Prentice Hall.

Lam, V. 2005. Causation and space-time. *History and Philosophy of the Life Sciences* 27(3–4): 465–478.

Norton, J.D. 2007. Causation as folk science. In *Causation, physics and the constitution of reality*, ed. H. Price and R. Corry. New York: Oxford University Press.

Rueger, A. 1998. Local theories of causation and the a posteriori identification of the causal relation. *Erkenntnis* 48: 25–38.

Russell, B. 1912. On the notion of cause. *Proceedings of the Aristotelian Society* 13: 1–26.

Salmon, W. 1997. Causality and explanation: A reply to two critiques. *Philosophy of Science* 64: 461–477.

Good Just Isn't Good Enough: Humean Chances and Boltzmannian Statistical Physics

Claus Beisbart

1 Introduction

Sometimes, just being good isn't good enough. Sometimes, you have to be better than your competitors. You have to be best, and, maybe, even by far more better than all other rivals. This, it seems, is not only true in sports, but also in metaphysics. For a probability to be a real-world chance, it has to be very good. It can't just be any old Bayesian degree of belief or so. It has to be best; in fact, it has to be much better than any other probabilities. This, at least, follows from David Lewis's account of chances. According to Lewis, chances are best probabilities, i.e., probabilities that occur in a best system.

The question of this paper is whether probabilities assumed in statistical physics are good enough to be chances. Are they real-world chances in the sense defined by Lewis?

Why ask this question? Well, the interpretation of probabilities[1] from statistical physics is a difficult problem. On the one hand, there are strong motives to take them to be ontic probabilities, i.e., probabilities that are independent of human needs and interests (Loewer 2001, pp. 611–612). On the other hand, probabilities used in classical statistical mechanics have to be compatible with determinism. But how can there be chances in a deterministic world (Loewer 2001)?

According to B. Loewer (2001, 2004), there can be chances in a deterministic world. The idea is that a certain probability distribution over the initial condition of a classical Universe is good enough to yield chances. In the terms used by Callender

[1] See Hájek (1997), pp. 210–211 and Hájek (2010) for the interpretation of probabilities quite generally.

C. Beisbart (✉)
Institute for Philosophy, University of Bern, Länggassstr. 49a, CH-3012 Bern, Switzerland
e-mail: Claus.Beisbart@philo.unibe.ch

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_36,

(2011), Loewer takes a globalist strategy. The details of Loewer's account have been criticized by Frigg (2008). Frigg and Hoefer (2010, forthcoming) have thus pursued a different strategy, which is localist in the terms of Callender.

I will argue that a globalist strategy runs into problems, but not quite for the reason that Frigg (2008) provides. I will then argue that a localist strategy, though more promising in one respect, doesn't do much better. The gist of my central objection is that good just isn't good enough. Winsberg (2008) is also critical of Frigg and Hoefer, but not quite for the reasons given in my paper.

The plan is as follows. In Sect. 2, I will shortly sketch the Humean approach to chances. Section 3 argues that macro probabilities concerning thermodynamic behavior may be part of the best system. Section 4.2 turns to statistical mechanics and examines the localist and the globalist strategies. Objections and replies are presented in Sect. 5. I draw my conclusions in Sect. 6.

This paper is subject to a few limitations. I will only consider classical statistical mechanics and assume that the world is deterministic (which I don't take to be true). My question really is whether probabilities that statistical physicists assume for a classical world are chances. Further, I focus on Boltzmannian statistical mechanics in the way presented in Albert (2000). My focus is on the explanation of how equilibrium is approached. I will not consider approaches that rely on ergodicity or a variety thereof (see Lavis (2011) and Frigg and Werndl (2011b) for new developments). Finally, as Meacham (2010), p. 1133 observes, there aren't yet many works that cash out the details of a Humean account of chances in statistical physics (with some exceptions, e.g. Frigg and Hoefer forthcoming). This paper will try to advance things a bit in this respect. Nevertheless, it is still sketchy and I offer my apologies for this. The only excuse is that this is more or less the state of the art.

2 How to Win the Race: A Short Guide to Humean Best System Probabilities

According to Lewis, probabilities are chances if and only if they are in some sense best. But what does it mean to be best and what are the standards of the competition? In this section, I point out how I understand Lewis's proposal, and I introduce a few assumptions that will be important for my argument.

For Lewis, probabilities are best if they are part of the best system. Roughly, the best system provides something like an optimal pocket guide to the Universe. It is very informative, but simple.

In more detail, the systems in the competition consist of sentences. Some sentences are non-probabilistic, others assign probabilities to events or propositions.[2] The corresponding probabilities are purely formal, they are not yet supposed to

[2]Probabilities are assumed to apply to propositions, but, for convenience, I will sometimes also speak of the probabilities of events.

refer to chances. As a consequence, the probabilistic sentences cannot be true or false. But otherwise, the competition is restricted to systems that are true in their non-probabilistic sentences.

The race is defined in terms of three pro-tanto desiderata. First, other things being equal, a system is better than another if it is *stronger*. In the simplest case, strength is just logical strength. But things are not always simple, in particular, logical strength does not apply to probabilistic sentences that lack definite truth values. Nevertheless, a system Σ is certainly stronger than another, Σ', if it entails probabilities for all propositions to which Σ' ascribes probabilities and if it entails probabilities for propositions for which Σ' does not entail probabilities. The idea is that Σ makes progress compared to Σ' because it provides at least probabilities for some propositions with respect to which Σ' does not say anything at all.

Second, other things being equal, a system is better than another if it is *simpler* (it's a pocket guide for tourists, so it can't be too complicated!). The appeal of simplicity is clear enough, but it's also clear that more has to be said about simplicity. In this paper, we need only compare the simplicity of what I call *equiprobability models*. These are probability models that are zero everywhere apart from a region D, in which the probability density is a multiple of the Lebesgue measure. D is called the support of the model. I assume that an equiprobability model is simpler than another if the support can be identified using less information.

Third, other things being equal, a system is better than another if it fits the patterns of the actual world better. For a very simple example, if a coin is flipped ten times and lands heads four times, then a probability of 0.5 fits the pattern of outcomes much better than a probability of 0.6, because the actual outcomes of the trials are more probable under the first probability model than under the second. More generally, a probability model is the better the more likely the patterns of the real world are under the model. This idea is very difficult to generalize to probability models that live on infinite spaces (Elga 2004), but this will not be a problem in what follows. I will only need an assumption about equiprobability models with support regions that each include all actual data points relevant for the model. One of these models yields better fit if its support is a proper subset of the support of the other. A possible motivation is that the first model provides more information than the other because the actual data points are more narrowly circumscribed.

The three criteria apply other things being equal, but often other things are not equal. Often, you can improve the fit of a probability model by making it more complicated. To obtain an overall comparison between both models and the corresponding systems, we have to strike a balance between the desiderata. Lewis assumes that, at least in many cases, there is only one rational way to do so. Overall, one system is better.

In this way, pairs of systems are compared. It is plausible to assume that there are limits to the quality of systems and that there is a best system. The probabilities in the best system win the race, and as a prize, they become chances.

But is it always possible to single out one winner? There may be two systems with two different probability models, but we may not be able to determine in a

unique rational way which is better. It then is not clear which system is better overall. The most promising strategy to deal with this problem seems to deny that there are any chances in this case. The idea is that we only speak of chances if the competition singles out a unique winner that is by far more better than the other systems such that there can be no quarrels as to who the winner is.[3]

There are complications and problems that we can bracket in what follows.

First, for the purposes of this paper, the exact nature of the pattern to which probabilities are fitted can be left open if only the pattern is non-probabilistic.

Second, in this paper, I will not draw on the so-called Principal Principle (PP, for short; Lewis 1980). The principle is not proper part of the best-system account of chances (although it may follow from it, see e.g. Hoefer 2007), and its precise formulation is controversial (see e.g. Lewis 1994).

Third, as Lewis himself is aware (Lewis 1994, p. 479), simplicity may ultimately turn out to be a mind-dependent. If this is so, chances are not really objective. I will leave this problem at one side.

Fourth, Lewis submits that chances are incompatible with determinism (Lewis 1986, pp. 117–121). But Loewer (2001, 2004), Hoefer (2007) and Frigg and Hoefer (2010) argue that indeterminism doesn't follow from the core of Lewis's account, but only if further commitments on the part of Lewis are taken into account. It is then suggested that these commitments should be given up. I will follow Loewer, Frigg and Hoefer in assuming that chances do not imply indeterminism.

Finally, Hoefer (2007) and Frigg and Hoefer (2010) ban non-probabilistic statements from their systems altogether and only consider probability models that are fitted to patterns in the real world. But non-probabilistic regularities do play an important role even if our main interest is in chances. For we do not assign chances to patterns that are fully described in terms of deterministic laws. Frigg and Hoefer have to rule this out by hand. Indeed, Hoefer (2007), pp. 563–564 requires that the patterns to be captured by chances look random. This requirement is not necessary if we allow for non-probabilistic statements in the systems.

3 Thermodynamics and Chances

Our main question is whether the probabilities assumed in statistical mechanics are chances in the sense defined by Lewis. Some of these probabilities are supposed to explain the validity of the so-called Second Law of Thermodynamics. For the remainder of this paper, it is useful to have a brief look at thermodynamics. In this section, I will suggest that the Second Law may be a statement about chances.

[3]This strategy is employed by Lewis concerning laws (Lewis 1994, p. 479). I'm not aware that he uses the same strategy concerning probabilities, but, in any case, it seems a very natural move. Cf. Winsberg (2008), p. 881.

This suggestion is absurd if the Second Law claims that entropy is always increasing until the equilibrium is attained. But such a claim would be too strong. In fact, micro-physics and statistical physics would not underwrite such a strong claim.

In a more cautious formulation, the second law reads (cf. Frigg 2008):

> SL_0 For all times t, if an isolated macroscopic system has an entropy $S(t)$ at time t, then, for every $t' > t$, the entropy $S(t') \geq S(t)$ with a very high probability.

We may add that, very likely, the entropy increases unless the system is already in equilibrium. To keep things simple, I bracket this part of the Second Law, but nevertheless talk of entropy increase.

In this version, the Second Law is about probabilities. The question thus arises whether the probabilities are chances?

SL_0 is only qualitative because no values of the probabilities are specified. Lewis's account of chances, by contrast, only applies to quantitative probabilities. But SL_0 may only be a first shot at a more precise, quantitative version of the Second Law. Here is how the more quantitative version may look like:

> SL For all $N \in \mathbb{N}$, for all energies E, for all times t, if an isolated macroscopic system of energy E with N degrees of freedom has an entropy $S(t)$ at time t, then, for every $t' > t$, the entropy $S(t') \geq S(t)$ with a probability of $P(N, E, S(t))$.

Here $P(N, E, S)$ is a real function with values in [0,1]. For most values of N, E and S, $P(N, E, S)$ is very close to one, but apart from this, the exact shape of the function is not yet known.

SL provides unique probabilities for the increase in entropy. The probabilities depend on the size of the system, its energy and its current entropy. Maybe, the function should also depend on other characteristics, e.g. on the strength of the most important interaction. It is not my task here to specify $P(N, E, S)$; future physics should try to do so.

The question then is whether $P(N, E, S)$ is a chance function. In a Lewisian framework, this is to ask whether $P(N, E, S)$ is good enough to be a chance function? To be more precise, is it possible to pick a function $P(N, E, S)$ such that it is part of the best system?

The best system in a classical world will certainly include the classical deterministic micro-dynamics. Call the system that only contains the deterministic micro-physical laws $\Sigma(CM)$. Add SL to $\Sigma(CM)$ and call the resulting system $\Sigma(CM, TD)$. If $\Sigma(CM, TD)$ is an improvement on $\Sigma(CM)$, and if it is so in an optimal way, as far as the description of entropy is concerned, then there is a strong reason to think that the best system includes $\Sigma(CM, TD)$, which is sufficient for $P(N, E, S)$ specifying chances.[4]

[4]Things are in fact more complicated. Even if $\Sigma(CM, TD)$ is an optimal improvement on $\Sigma(CM)$, as far as entropy is concerned, we may be able to move to an even better system that covers other aspects of the pattern and entails a different probability function for the increase of entropy. But this possibility will not matter for the purposes of my argument.

To check whether the addition of SL improves on $\Sigma(CM)$ in an optimal way, we have to apply our criteria.

- Strength: $\Sigma(CM, TD)$ is stronger than $\Sigma(CM)$ because a system is stronger than another if the Boolean algebra of propositions to which probabilities are assigned becomes larger. Admittedly, we would obtain at least as much strength if we added a non-probabilistic statement about the values of entropies in isolated systems. However, assuming that there are no deterministic laws about the increase of entropy, this statement would either be false or extremely complicated. Both ways it could not be part of the best system.
- Simplicity and fit: $\Sigma(CM, TD)$ is less simple than $\Sigma(CM)$ because we have added a sentence to $\Sigma(CM)$ that is unconnected to others. In particular, it uses two predicates that do not occur otherwise in the system, viz. "isolated system" and "entropy".[5] However, at least the predicate "isolated system" can easily be defined in terms of the micro-physics (the condition being that there are no interactions with other systems). Further, maybe, even entropy can be defined in terms of the micro-physics. In any case, the loss in simplicity is not huge if the form of $P(N, E, S)$ is relatively simple.
  Now we don't know this form. As a consequence, neither simplicity nor fit can really be specified. But it seems at least possible that there is one unique model that provides reasonable fit, but is sufficiently simple and thus strikes a the unique rational balance between both desiderata.

If there is such a function, then adding it to $\Sigma(CM)$ provides an optimal improvement on $\Sigma(CM)$, and $P(N, E, S)$ specifies chances.

4 Probabilities in Boltzmannian Statistical Physics

4.1 Boltzmannian Statistical Physics

Boltzmannian statistical physics tries to explain the Second Law in the following way. Each isolated macroscopic system is assumed to consist of a great number (N) of micro-constituents, typically atoms or molecules. The dynamics of the system is described using a $6N$-dimensional phase space. Since the system is isolated, energy is conserved, and the motion is confined to a $(6N-1)$-dimensional hyperplane Γ in phase space. Each possible micro-state of the system corresponds to a point on the hyperplane; as time passes, the systems takes a trajectory in Γ.

From the macroscopic point of view, the system is described in a coarse-grained way only using macro-information, e.g. state variables such as temperature. Using

[5]Note that the word "system" is ambiguous in this paper, it either refers to systems of representations (e.g. Lewis's best system) or to physical systems. I assume that readers will always be able to find out which sort of systems I refer to.

the macro-information, we can define macro-states M_i. Each macro-state M_i can be realized by a great number of micro-states and thus corresponds to a subset of Γ, call it Γ_i. Altogether, the macro-states induce a partition of the energy hyperplane Γ.

The phase-space volume of a macro-state can be measured in a very natural way using the $(6n-1)$-dimensional Lebesgue measure μ over Γ. The entropy of macro-state i is defined as

$$S_B(M_i) = k \ln(\mu(\Gamma_i)), \tag{1}$$

where k is the Boltzmann constant. The entropy of the system at time t is then defined via the unique macro-state $M_{i(t)}$ in which the system is at t:

$$S_B(t) = k \ln(\mu(\Gamma_{i(t)})). \tag{2}$$

At least for simple systems, this entropy can be shown to coincide with thermodynamic entropy (see Frigg and Werndl (2011a) for details). In the following, I will always assume that this equality holds. I will thus drop the subscript "B" in the following.

The main idea then is to introduce a probability distribution over the micro-states and to derive the proposition that entropy increases with a very high probability. Roughly, the crucial probability model locates the system in its current macro-state and is flat or homogeneous within this macro-state. That is, within this macro-state, the probability measure follows the Lebesgue measure. Otherwise, it is zero. The region over which a non-zero probability density is assumed is called the support of the probability model. The equilibrium state has a maximal phase space volume, and it is intuitive that most trajectories end up being there for a long time.

According to current wisdom, the story can only work if we additionally assume the Past Hypothesis, viz. that the initial state had a very low entropy (see Albert (2000) for an influential formulation). The trick is most often applied to the whole Universe, but we can also apply it just to the system under consideration (more on this below). In the remainder of this section, we will apply this strategy to an isolated system, be it the Universe or not.

Here are the details (I orient myself after Frigg 2008): let t_0 be the initial time. For each macro-state i and each time interval Δt, define a region in phase space $B_{i,\Delta t}$ as follows. $B_{i,\Delta t}$ comprises all those points within Γ_i that evolve, within time interval Δt, into states corresponding to an entropy at least as high as $S(M_i)$.[6] Denote the initial low-entropy macro-state by M_{i_0}. Let Γ_t^0 be that phase space region in which trajectories beginning in the Past State end up at t (the following would also work if the Past State were defined as a union of macro-states) and consider the following claim.

[6]It is here assumed that the dynamics is homogeneous or stationary in time, i.e., if $\mathbf{x}(t)$ is a possible trajectory in phase space, so is $\mathbf{x}(t+T)$ with a constant T.

PD Phase space dynamics: for each $\Delta t > 0$ for each $t > t_0$, for each i: $\mu(B_{i,\Delta t} \cap \Gamma_t^0) \approx \mu(\Gamma_i \cap \Gamma_t^0)$.

PD states that most trajectories that start from the Past State and that are in a specific macro-state i at t run into regions of phase space with an entropy at least as high as that of i. Here, trajectories are "counted" using the Lebesgue measure. PD is an assumption about the dynamics. If PD or some suitable variant thereof holds true, it should follow from the micro-dynamics of the system. In this paper, I will take it that PD or a suitable variant thereof can be shown.

In addition, we need an assumption about probabilities over micro-states.

E For each time $t > t_0$, the probability that the system is in phase space region B is

$$\frac{\mu(B \cap \Gamma_t^0 \cap \Gamma_{i(t)})}{\mu(\Gamma_t^0 \cap \Gamma_{i(t)})}. \tag{3}$$

$M_{i(t)}$ is the current macro-state, $\Gamma_{i(t)}$ the phase space region corresponding to it. E is an equiprobability assumption because it assumes probability densities that are homogeneous (or flat) concerning the Lebesgue measure. The homogeneous distribution extends over that part of the region of the current micro-state that can be reached from the Past State.

PD and E imply that, for every t and for every Δt, there is a high probability, that $S(t + \Delta t) \geq S(t)$. Thus, a qualitative version of the Second Law follows. Ideally, we can strengthen PD and E to obtain SL, but this will not be our concern in what follows. There may also be other improvements on our formulations of SL, PD and E. I do not think that this affects my arguments concerning probabilities.

E introduces probabilities over micro-states. Our question is whether they are chances. To answer this question, we have to distinguish between two strategies. Both underwrite E, but in different ways. Globalists assume a low-entropy state for the whole universe, assume a probability distribution over the micro-state and try to derive everything else. Localists postulate a low-entropy state and a certain probability distribution for each thermodynamic system (see e.g. Callender (2011) for the distinction). In the following, I will consider each strategy in turn.

The setting is always as follows. We start with the system $\Sigma(CM, TD)$ and add certain probabilistic assumptions that lead to E. The question is whether the assumptions improve on $\Sigma(CM, TD)$ in an optimal way. If so, there is a good case that they form part of the best system of the world and that the related probabilities are chances.

4.2 Globalism About Micro-probabilities

Turn first to globalism as adopted by Loewer (2001, 2004). Granted the deterministic laws that fully describe the micro-dynamics of the world, we would gain a lot of information if we added sentences fully describing the initial condition. However, we will also pay a lot in simplicity, because information about the initial condition

has to take the form of a myriad of unconnected sentences (there is one particle at position x and with momentum p, and so on). Adding precise information about the initial condition will thus not improve the system overall. But there seems to be an alternative: we fit a simple probability distribution to the initial condition.

But how can we fit the initial condition using a simple model? According to Loewer (see in particular the reconstruction by Frigg 2008), an equiprobability distribution is assumed within the region corresponding to the initial macro-state, which has a very low entropy according to the Past Hypothesis. Roughly, micro-states not compatible with the macro-information have zero probability of occurring, while micro-states within the right part of the phase space are equally likely.

Adding this probability model over the initial condition clearly adds strength. The model is also simple and fit seems reasonable. So we have made some progress by adding the model to $\Sigma(CM, TD)$. But maybe there is an alternative way of improving upon $\Sigma(CM, TD)$?

We can improve fit when we further shrink down the support of the probability model. That is, we assume a flat probability distribution over a certain sub-region of the past low-entropy macro-state. That sub-region may be defined by the demand that a certain elementary particle has a kinetic energy larger than a particular value e_0, for instance.

If we do so, we have to pay in simplicity though because the new probability model is less simple than the old one. The new probability model is still an equiprobability model, but it is less simple to characterize its support. To characterize the support, we begin with the support of the old probability model and further demand that the kinetic energy of a particular particle be e_0.

So overall, would we improve the system by shrinking down the support? This is a difficult question, and the answer is far from clear. The only thing we can say is that fit is considerably improved, but that there is a considerable cost in simplicity too. So it's not a case in which the right sort of balance favors one system rather than the other in a clear way.

But this is enough to conclude that the equiprobability model over the Past State can't provide chances. For as you will recall, good just isn't good enough for a best system. We need a system that is clearly optimal, but Loewer's isn't really. The consequence is that E as applied to the whole Universe isn't part of the best system. The probabilities aren't chances.

It may be objected that we do not have any information about the value of the kinetic energy of a particular particle. But this is immaterial for our purposes. Our ignorance only implies that we are not in a position to know the initial probability distribution. But our interest is in metaphysics, not in epistemology. The question is what the unique probability distribution over initial condition is. And for fixing this distribution, human needs and limitations do not matter (as Winsberg 2008 argues forcefully).

Further, under certain assumptions, there are two probability models that do not require the evaluation of micro-information, but that are still roughly equally good. For suppose that there are two macro-states with the same small phase-space volume and thus with the same small entropy. Suppose further that the initial condition of

the Universe is in one of the two macro-states. There are the two following options for a probability model: either we assume a flat probability model over the actual past macro-state or we assume a flat probability model over the set union of both macro-states. The first model provides better fit, but is more complicated because the support region is more difficult to pick: it is not sufficient to demand that entropy be minimal, rather we also have to use additional macro-information to pick the actual Past State. The second model does not fit as well as the first, but it is simpler. Again, there is no clear winner between the models. Both improve on the system $\Sigma(CM, TD)$, but we cannot tell which model provides a better improvement.[7]

Frigg (2008), p. 679 (cf. also Frigg and Hoefer forthcoming, Sect. 5.5) seems to think that there is in fact a probability model that is clearly optimal, viz. a Dirac delta distribution peaked at the actual initial condition. So, for each state $\mathbf{x}$, the chance distribution would be

$$p(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_0), \tag{4}$$

where $\mathbf{x}_0$ is the $(6N - 1)$-dimensional vector specifying the actual initial condition, i.e. the positions and momentums of all particles at time t_0. Clearly, the delta distribution provides optimal fit. Frigg further thinks that the delta distribution is still simple because it has a simple functional form. Overall, he concludes, the delta distribution is optimal.

However, in my view, Frigg overlooks a lot of costs. The delta distribution has a simple functional form, but to pick one particular delta function, you have to specify the location of the peak, $\mathbf{x}_0$, i.e., the whole initial condition, and this is not simple at all![8]

The globalist approach runs into other troubles too. So far, we have only probabilities over the initial condition. But E refers to probabilities at later times. So we need introduce a dynamics of the probabilities.

There is a very natural way to fix a dynamics. We let the chances follow the micro-dynamics. Consider the probability that the system is within region $B \subset \Gamma$ at time t. The probability of finding the system in B then is assumed to be

$$\mu(B \cap \Gamma_t^0)/\mu(\Gamma_t^0). \tag{5}$$

where Γ_t is the dynamical development of the phase-space part corresponding to the Past State.

But this dynamics does not underwrite E. In E, the probability is additionally conditioned on the macro-state in which the system is at time t. There are thus at least two alternative dynamics of the probabilities.

---

[7] Alternatively, we can shrink the support of the probability density over the initial condition by taking into account macro-information about later times, e.g. by conditioning on the macro-state 1 s after the initial time (cf. Frigg 2008, p. 679). Once again, we gain fit, but pay in simplicity.

[8] I'm here in agreement with Meacham (2010), p. 1129 and Frisch (2011), pp. 1004–1005.

From a Humean perspective, chances follow the dynamics that leads to the best system. We need the dynamics that strikes an optimal balance between simplicity and fit (strength is not an issue because both alternative dynamics are defined on the same Boolean algebra). Conditionalization on the current macro-state improves fit because we narrow down the range of possible micro-states. However, we have to pay in simplicity. The formula for the probability distribution itself becomes more complicated due to the conditionalization. Is there a clear balance between simplicity and fit that decides between both probability models? I do not think so. Again, we are faced with a situation in which there is no unique rational balance between simplicity and fit. There is no unique winner. This means that both probability distributions do not specify chances.[9]

To be sure, if you have to guess what the real micro-state of the system is, then you do much better if you conditionalize on the present macro-state because you then take into account information that is pertinent to the case at hand. But we are here not guessing what the present micro-state is. We are in the business of finding a dynamics of chances that produces a best system. Better guesses go along with better fit, but fit is only one item on the list of desiderata.

Loewer goes even further and suggests that we condition on the whole macro *history* of the system under consideration. The effect is that fit is further improved while we have to pay in simplicity. Further, why not condition on further characteristics?

The conclusion looming large here is that Humean considerations do not provide us with any clearly optimal solution to the problem of how to define a dynamics of chances. And good just isn't good enough. We need a best system that is clearly best, and not just a good one. If there isn't a best system, then there are no chances, or so I have suggested above. This means that the probabilities introduced by the globalist strategies cannot be interpreted as chances according to Lewis.

There is a further problem with the globalist strategy to underwrite E (e.g. Callender 2011, pp. 100–102). The account justifies E as applied to the whole Universe (the isolated system being the whole Universe). But what we are to explain is not only a global increase in entropy, but rather local increase in entropy in a lot of systems, e.g. in coffee cups together with their environments. Thus, the globalist approach only makes sense if E as applied to the Universe implies the right kind of probabilities about small systems. This is a very long way to go. Entropy is additive, and that global entropy increases is no more than to say that entropy is increasing on average. This leaves space for large pockets of decreasing entropy.

[9]Note that our natural dynamics of chances does not really mesh with PD. To explain the increase in entropy, a modification of PD would have to be combined with the natural dynamics.

4.3 Localism About Micro-probabilities

Turn now to the localist strategy as taken by Frigg and Hoefer (forthcoming).[10] This is a different attempt to improve on the system $\Sigma(CM, TD)$. The crucial move here is to apply E to all isolated thermodynamic systems. Here is the precise localist formulation of E:

> E_l For each system within the class of isolated thermodynamic systems, for each time $t > t_0$, the probability that the system is in region B is
>
> $$\frac{\mu(B \cap \Gamma_t^0 \cap \Gamma_{i(t)})}{\mu(\Gamma_t^0 \cap \Gamma_{i(t)})}. \tag{6}$$

The initial time t_0, phase space and the respective partition into macro-states are now defined for each system s separately (t_0 may that time at which the system is isolated from its environment). Note that we are implicitly quantifying over phase spaces of various dimensions because the systems under consideration differ in their numbers of degrees of freedom. Further, the systems will differ in their initial states, so we have to quantify over Γ^0 as well. Note finally that E_l quantifies over instances of time and thus fixes a probability distribution for every time. There is no need to postulate a dynamics of chances on top of this; the dynamics of chances is basically given by the micro-canonical distribution under the constraint of a low initial entropy.

E_l is a very clever way to systematize information about patterns found in the world. The idea is that there are patterns over and above the deterministic laws of succession, and these patterns arise because certain types systems arise in history and display a certain recognizable behavior. Ultimately, this is due to the initial condition of the Universe and the dynamics, but, even granted the deterministic laws, the patterns enfolding in history can't be translated to a simple pattern in the initial condition.[11]

The proposal then is that E_l is a part of the best system. As Frigg and Hoefer (forthcoming), Sect. 5.5 note, from a Humean perspective, there is something healthy about E_l. E_l claims there to be a regularity and is thus often instantiated. The probability distribution that E_l claims for every isolated system can thus be obtained using a myriad of systems of the same type. The hope would be that this suffices to obtain a unique probability distribution by accounting for the relevant patterns in the Humean way.

[10] In this section, I focus on the localist strategy based upon the work of Albert. It closely parallels the globalist strategy as outlined above. Frigg and Hoefer (forthcoming) take the localist strategy in a different way. But they run into problems parallel to those outlined in this section. See footnote 14 below for details.

[11] The same point applies to SL probabilities.

But do we really improve on $\Sigma(CM, TD)$ in an optimal way if we add $\mathrm{E_l}$ to it? Let us take the criteria in turn.

- Strength: We certainly gain strength because the Boolean algebra over which probabilities are defined is extended.
- Simplicity: The probability distribution over micro-states is fairly simple.
- Fit: E quantifies over phase-space dimensions and thus over probability models. To determine fit, we have to take each phase space in turn. We begin, maybe, with a phase space of $6N$ dimensions for some N and consider all possible energies E, continue with the next of $6(N+1)$ dimensions and so on. For each phase space and energy E, we consider the entirety of systems that can be characterized using the hyperplane with energy E in this phase space. Depending on how large the Universe is and on what the distribution of the micro-systems is like, we may only obtain very few systems per phase space and energy E (if any at all) or a large number.

  How good will fit be? This is a good question. Nobody has yet done the exercise to check whether systems with a $6N$-dimensional phase and with energy E scatter homogeneously within the range defined by the macro-variables or not, and the check is a difficult exercise too. So the correct answer seems to be that we don't know (cf. Frigg and Hoefer forthcoming, Sect. 4 for this point).

Altogether, since we don't know what fit is like, we can't really tell whether it pays off to add $\mathrm{E_l}$ to the system $\Sigma(CM, TD)$. So we can't really tell whether the probabilities are chances.

But suppose now for the sake of argument that the equiprobability models in the phase spaces provide good fit such that $\mathrm{E_l}$ leads to an improvement on $\Sigma(CM, TD)$. Even then, we face a problem. Consider the following alternative to $\mathrm{E_l}$,

> $\mathrm{E_l'}$ For each system from the class of isolated systems, for each time $t > t_0$, the probability that the system is in region B is
>
> $$\frac{\mu(B \cap \Gamma_t^0)}{\mu(\Gamma_t^0)}. \tag{7}$$

$\mathrm{E_l'}$ does not take into account the current macro-state and doesn't delimit the equiprobability distribution to the corresponding part of phase space. In this way, fit becomes worse, while we gain in simplicity because the distribution is now less complicated (the model has less parameters). Is it better to add $\mathrm{E_l'}$ or $\mathrm{E_l}$ to $\Sigma(CM, TD)$? This is a difficult question. It doesn't seem to be a clear case in which the balance between simplicity and fit favors one side without any doubts.

Maybe, we can even go the other way round and move to $\mathrm{E_l''}$ that conditionalizes on additional information (be it macro or not). Fit can be improved considerably in this way, but we have to pay concerning simplicity. If we strike the balance, it is not clear which hypothesis, $\mathrm{E_l''}$ or $\mathrm{E_l}$, leads to a better system. And, recall, good just isn't

good enough. So we haven't found an optimal system, and if there isn't, then there aren't Humean chances. In conclusion, both the globalist and the localist strategies don't yield chances over micro-states.[12]

5 Discussion Points

Let me now discuss the argument and the results of this paper.

A natural reaction to them is as follows. Lewis's account of chances, it may be said, does not work quite generally. The competition for the best probabilities can't ever produce a winner. It's simply not possible to weigh simplicity against fit in a unique rational manner, in statistical physics as everywhere else. Or so it may be suggested.

I don't think that this is true. Lewis's account describes quite well what scientists do when they use probabilities to describe phenomena in nature. Sometimes they do come up with unique probabilities the values of which are not much debated. In these cases, Lewis's account should work. As an example, consider electrons with a spin parallel to a certain direction, call it z. Spin is later measured in a direction perpendicular to the z-direction, call it y. We restrict our attention to systems in which no forces affect spin between the preparation of an electron and the measurement. Consider now the outcomes of all these experiments. The outcomes define what we may call an empirical distribution. As is well-known, there is no way to predict the outcome of an individual measurement on the basis of information about the experimental set-up. Consequently, the outcomes cannot be inferred from initial conditions and laws of nature.[13] To include information about the outcomes in a system, we may compose a large list of all the individual outcomes, but this would render the system too complicated. The only way left to cover the outcomes is thus to assume a probability distribution. We fit a simple probability model to the empirical distribution. In our examples, the model has it that an electron has a probability of 50 % to have a spin of $+1/2$ in the y-direction.

The probabilities under consideration are chances as defined by Lewis if there is no alternative probability model that does at least roughly equally well on the desiderata. Let us thus try to find an alternative model that is roughly as good as our first one. We could increase fit, if we partitioned the electrons into two broad classes that displayed markedly different empirical distributions over spin measurements in the y-direction. We would then fit different probability models to the empirical distributions from both classes. But we would have to pay in simplicity because having different probability models for the several sub-classes makes things more

[12]To be fair to Frigg and Hoefer (forthcoming), we should note that they explicitly bracket the question as to whether their strategy yields unique best probabilities (Sect. 4).

[13]Bohmians, of course, deny this, but they nevertheless grant quantum-mechanical probabilities.

complicated. All in all, it might prove difficult to strike a balance between simplicity and fit, and we would end up in the same sort of mess that we encountered in statistical physics.

But can we indeed pick two classes with markedly different empirical distributions of spins? In principle, we can, we may for instance pick by hand those electrons with a positive y-spin. However, in this way we would lose a lot of simplicity. The reason is that a system with two probability models over spins has to specify to which class of electrons a probability model pertains. If we pick the classes by picking individual electrons by hand, as it were, the resulting system becomes too complicated. It will clearly be worse than the system with one probability model over all spins.

But maybe we can find an alternative system that partitions the outcomes of the spin measurements in a much simpler way and nevertheless leads to two markedly different empirical distributions. The only way to do this is to pick the classes in terms of general physical characteristics. For instance, we may say that the first class includes those electrons with kinetic energies higher than e_0, whereas the second class comprises the other electrons. The crucial problem now is that there is no such way to partition the electrons in two classes with markedly different empirical distributions. At least as far as we know, there is no physical characteristic that correlates with spin in these kinds of experiments. In whatever way we partition the class using natural predicates built up from well-known physical characteristics, we always find roughly the same empirical distribution over spins. The distribution is invariant under any natural partition (under any partition defined using well-known physical characteristics such as mass ...). There is thus no way to improve fit by paying a bit of simplicity. Subsuming all electron spins under one general statement is the best way to account for them from the Humean perspective.

Things are different in the case of isolated systems, over which we quantify in statistical physics. There are always well-known natural variables that allow us to improve fit. The reason is that our systems are built up of micro-constituents and that we can use information about the latter to improve fit. This leaves us with a number of systems that are about as good as the others. And good just isn't good enough, as you will remember by now.

The example in which we do have chances according to Lewis is from quantum mechanics, which may be thought to be indeterministic. But what about probabilities for macro-systems in a world with a deterministic micro-dynamics? Can Lewis's account ever produce chances at this level? A worry might be that it cannot, and if it cannot, then Lewis's account may seem implausible.

But I think, it is indeed possible that Lewis's account yields chances even in this case. The account yields chances if a certain type of macro-event (e.g. that a coin lands heads) arises because a large number of degrees of freedom interact in a very complicated way, such that the empirical distribution over the macro-events is invariant or almost invariant under partitions formed using simple combinations of physical characteristics.

This condition may be fulfilled for thermodynamic behavior. Consider the class of thermodynamic systems with a fixed number of degrees of freedom.

There are a few parameters (energy, entropy, strength of interaction, density) that systematically influence the empirical distribution over increase in entropy. But once these parameters have been identified and taken into account in a model $P(N, E, S, \ldots)$, it seems, there are no simple ways to subdivide the class further to improve fit. Part of the reason would be that increase in entropy is multiply realizable in a great many ways. If all this is correct, then there is a unique best probability model, and we do obtain chances. The question of whether the right sort of conditions obtain for there being chances over entropy increase cannot be answered in this paper, it would require more physics.

In any case, in statistical physics, things are different. In thermodynamics, we are interested in the probability of macro-events. The macro-events are multiply realizable, so there is an open question as to whether we could improve fit by defining suitable sub-classes. By contrast, it is essential for statistical physics to specify probabilities over micro-events, and statistical physicists always assume a homogeneous probability distribution at some level. They can always improve fit by playing around with the support of the probability model. As fit becomes better, costs in simplicity arise, and it is not clear what the overall best model is.

So far, I have considered the Boltzmannian account of the Second Law. I have argued that the equiprobability models that are assumed cannot be taken to be chance distributions. But maybe the Boltzmannian account can be improved on. Maybe, there are actually much better probabilities that do in fact form part of a clear winner among the possible systems.

Here is one idea (see e.g. Winsberg 2008, Sect. 2; cf. also Frigg and Hoefer forthcoming, Sect. 3). Instead of introducing a flat probability distribution in some part of phase space, we rather partition each phase space into two regions, viz. one with initial conditions that lead to the right kind of thermodynamic behavior, and the other with initial conditions that do not. We take the relative frequencies or some elaboration of them from our pattern and identify them with the respective probabilities. These probabilities are then used to explain why thermodynamic behavior is very common.

However, this strategy gives up the project of statistical physics as it is commonly conceived of because we are only left with time-less probabilities over certain histories. Statistical physics, by contrast, seems to be committed to probabilities that a system is in a certain micro-state at a specific time. Further, the explanatory value of the probabilities would be small. It is clear that, in a deterministic universe, probabilities over certain events can be translated into probabilities over prior initial conditions. We gain not much by explaining the former in terms of the latter.[14]

Some people may want to suggest that the problem is with Lewis's account of chances and not with the probabilities in Boltzmannian statistical physics. Maybe,

[14]For an alternative, localist strategy, Frigg and Hoefer (forthcoming) do not just partition a suitable region of each phase space in two regions; rather, they assume a flat probability distribution over the initial macro-state in each phase-space. This then is supposed to imply a very high probability for trajectories that manifest thermo-dynamic behavior. But this suggestion runs into the same types of problems we have seen before. For instance, as Frigg and Hoefer (forthcoming), Sect. 4 themselves acknowledge, it is not known whether a flat distribution over micro-states yields good fit.

there are indeed chances, but Lewis's account fails to account for them. I don't think that this is a helpful suggestion. Lewis's account is fairly weak, and something like his requirements seem necessary for chances. The argument in this paper is that even these necessary conditions are not fulfilled.

6 Conclusions

In a nutshell, the argument of this paper runs as follows:

1. Statistical physics is built upon some variant of the equiprobability assumption E.
2. If the probabilities from statistical physics are best system probabilities, then E or a variant thereof has to express probabilistic assumptions that optimally improve on systems with deterministic laws and thermodynamics only ($\Sigma(CM)$ or $\Sigma(CM, TD)$).
3. Under both the globalist and the localist strategy, the pertinent probabilistic assumptions do not improve on $\Sigma(CM, TD)$ in an optimal way because there are rivals that are roughly equally good.
4. Thus, the probabilities in statistical physics are not best system probabilities or chances.

To conclude, let me put my argument in a broader perspective.

First, it has sometimes been argued that Lewis's account doesn't really provide ontic chances, either as a general view of probabilities or as applied to probabilities in statistical physics. Maybe, as Lewis himself fears, the notion of simplicity will ultimately have to refer to humans in one or the other way. Or, maybe, Lewis's account works only regarding statistical physics if we restrict any admissible information to macro-information, which would in the end introduce the perspective of human beings (see Winsberg 2008). My argument is not of this type. My argument may even have force if the task is only to provide an objective interpretation of probabilities in the sense of epistemic objectivity. For the argument denies that the notion of a best system suffices to fix the values of the probabilities uniquely, which would be essential for epistemic objectivity. In some cases, lack of knowledge or other epistemic constraints may lead to one unique probability model, for instance if we don't have any knowledge about micro-conditions. But this need not always be so.

Second, in this paper, I have only considered the approach to equilibrium. Probabilities from equiprobability models are used for other purposes too, for instance to derive relationships between state variables (pressure, temperature, entropy) in equilibrium. It may be argued that, for these purposes, E is essential and that we cannot change the probabilities in the way suggested in this paper. A possible reply would be that there are alternative probability distributions that are roughly on par with the usual ones, that produce the same macro-results (including the Second Law), but that do roughly as well as the usual ones in the best system approach. The idea would be that the support region in phase space can be narrowed down even further using other physical variables and that this doesn't affect the macro results (cf. Uffink 2011, p. 49 for a related remark). But to show this is left to future work.

Third, let me address as final worry. As suggested in the introduction, there is some pressure to defend ontic probabilities in statistical physics, and most ontic interpretations of probabilities do not mesh well with statistical physics. How then can we understand probabilities in statistical physics?

The point may be pressed even more. According to my argument, probabilities used in statistical physics don't belong to an uncontroversially best system because other probability functions would do as well. But statistical physicists do use certain probabilities and there doesn't seem to be much doubt about this. They do assume the micro-canonical distribution if they consider an isolated system. How can I account for this?

As an answer I'd like to suggest that, at least for many purposes, different probability distributions would do as well. I think, it doesn't really matter for the explanation of the approach to equilibrium how exactly we condition on macro-information that we happen to have. The reason is that arguments in statistical physics are robust, the details don't really matter (see Uffink 2011, p. 49 for a similar suggestion). There may be several probability models that strike a good balance between simplicity and fit, but they would all underwrite the approach to equilibrium. I haven't shown this rigorously, but this seems to be a fruitful perspective on this problem. The idea would be that statistical mechanics explanations are robust and that a lot of details don't matter.

Callender (2001) has rightly reminded us that we shouldn't take thermodynamics too seriously. Maybe we shouldn't take certain distributions from statistical physics too seriously, either. Put succinctly, my point then is this: Good just isn't good enough, if we consider the best systems approach, but it's good enough for statistical physics.

Acknowledgements I'm grateful to my commentators Luke Glynn, Radin Dardashti, Karim P. Y. Thebault and Mathias Frisch, to Georg Brun and to the participants at the Lausanne workshop for discussion. Thanks also to Michael Esfeld for the invitation and to Roman Frigg for sharing a yet unpublished manuscript with me.

References

Albert, D. 2000. *Time and chance*. Cambridge: Harvard University Press.

Callender, C. 2001. Taking thermodynamics too seriously. *Studies in History and Philosophy of Science Part B* 32(4): 539–553.

Callender, C. 2011. The past history of molecules. In *Probabilities in physics*, ed. C. Beisbart and S. Hartmann, 83–113. Oxford: Oxford University Press.

Elga, A. 2004. Infinitesimal chances and the laws of nature. *Australasian Journal of Philosophy* 82: 67–76.

Frigg, R. 2008. Chance in Boltzmannian statistical mechanics. *Philosophy of Science* 75(5): 670–681.

Frigg, R., and C. Hoefer. 2010. Determinism and chance from a Humean perspective. In *The present situation in the philosophy of science*, ed. D. Dieks, W.J. Gonzalez, S. Hartmann, M. Weber, F. Stadler, and T. Uebel, 351–371. Berlin/New York: Springer.

Frigg, R., and C. Hoefer. forthcoming. *The best Humean system for statistical mechanics*, forthcoming in Erkenntnis, doi: 10.1007/s10670-013-9541-5.

Frigg, R., and C. Werndl. 2011a. Entropy – a guide for the perplexed. In *Probabilities in physics*, ed. C. Beisbart and S. Hartmann, 115–142. Oxford: Oxford University Press.

Frigg, R., and C. Werndl. 2011b. Explaining thermodynamic-like behavior in terms of Epsilon-ergodicity. *Philosophy of Science* 78(4): 628–652.

Frisch, M. 2011. From Arbuthnot to Boltzmann: The past hypothesis, the best system, and the special sciences. *Philosophy of Science* 78(5): 1001–1011.

Hájek, A. 1997. 'Mises Redux'– redux. Fifteen arguments against finite frequentism. *Erkenntnis* 45: 209–227.

Hájek, A. 2010. Interpretations of probability. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta, spring 2010 ed. http://plato.stanford.edu/archives/spr2010/entries/probability-interpret/.

Hoefer, C. 2007. The third way on objective probability: A sceptic's guide to objective chance. *Mind* 116: 549–596.

Lavis, D.A. 2011. An objectivist account of probabilities in statistical mechanics. In *Probabilities in physics*, ed. C. Beisbart and S. Hartmann, 51–81. Oxford: Oxford University Press.

Lewis, D. 1980. A subjectivist's guide to objective chance. In *Studies in inductive logic and probability*, vol. II, ed. R.C. Jeffrey, 84–113. Berkeley: University of California Press. Here quoted from the reprint in Lewis (1986).

Lewis, D. 1986. Postscripts to 'A subjectivist's guide to objective chance'. In *Philosophical papers*, vol. II, ed. D. Lewis, 114–132. New York: Oxford University Press.

Lewis, D. 1994. Humean supervenience debugged. *Mind* 103: 473–490. Reprinted in Lewis, D. 1999. *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.

Loewer, B. 2001. Determinism and chance. *Studies in History and Philosophy of Modern Physics* 32: 609–620.

Loewer, B. 2004. David Lewis's Humean theory of objective chance. *Philosophy of Science* 71: 1115–1125.

Meacham, C.J.G. 2010. Contemporary approaches to statistical mechanical probabilities: A critical commentary – part ii: The regularity approach. *Philosophy Compass* 5(12): 1127–1136.

Uffink, J. 2011. Subjective probability and statistical physics. In *Probabilities in physics*, ed. C. Beisbart and S. Hartmann, 25–50. Oxford: Oxford University Press.

Winsberg, E. 2008. Laws and chances in statistical mechanics. *Studies in History and Philosophy of Science Part B* 39(4): 872–888.

Unsharp Humean Chances in Statistical Physics: A Reply to Beisbart

Radin Dardashti, Luke Glynn, Karim Thébault, and Mathias Frisch

1 Introduction

In an illuminating paper, Beisbart (2014) argues that the recently-popular thesis that the probabilities of statistical mechanics (SM) can function as Best System chances runs into a serious obstacle: there is no one axiomatization of SM that is *robustly best*, as judged by the theoretical virtues of simplicity, strength, and fit. Beisbart takes this "no clear winner" result to imply that the probabilities yielded by the competing axiomatizations simply fail to count as Best System chances. In this reply, we express sympathy for the "no clear winner" thesis, however we argue that an importantly different moral should be drawn from this. We contend that the implication for Humean chances of there being no uniquely best axiomatization of SM is not that *there are no SM chances*, but rather that *SM chances fail to be sharp*.

In Sect. 2 we outline the Humean Best System Analysis (BSA) of chance. In Sect. 3, we explain why it has been thought that the BSA justifies an interpretation of the SM probabilities as genuine chances. In Sect. 4, we describe Beisbart's arguments for the no clear winner result. As noted above, after establishing this

R. Dardashti (✉) • K. Thébault
Munich Center for Mathematical Philosophy (MCMP), Ludwig-Maximilians-University of Munich, Ludwigstr. 31, 80539 Munich, Germany
e-mail: Radin.Dardashti@lrz.uni-muenchen.de; Karim.Thebault@lrz.uni-muenchen.de

L. Glynn
Department of Philosophy, University College London, Gower Street, WC1E 6BT London, United Kingdom
e-mail: l.glynn@ucl.ac.uk

M. Frisch
Department of Philosophy, University of Maryland, 1108B Skinner Building, 20742 College Park, MD, USA
e-mail: mfrisch@umd.edu

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_37,

thesis, Beisbart goes on to conclude that there are in fact no SM chances. It is this second step that we wish to question, and we explain why by appeal to the notion of imprecise chances in Sect. 5.

2 The Humean Best System Analysis (BSA) of Chance

According to Humean theories of objective chance, the chances supervene upon the *Humean mosaic*: that is, the distribution of categorical (i.e. non-modal) properties throughout all of space-time. The most promising attempt to capture the manner in which the chances supervene on the mosaic is the Best Systems Analysis (BSA), which has received its most significant development by Lewis (1983, 1994).

According to the BSA, the objective chances are those probabilities that are entailed by that set of axioms which best systematizes the Humean mosaic, where goodness of systematization is judged against the theoretical virtues of simplicity, strength, and fit. The BSA is offered as an analysis of laws as well as of chances: the laws are the (axioms and) theorems of the Best System.

A system is *strong* to the extent that it says "what will happen or what the chances will be when situations of a certain kind arise" (Lewis 1994, p. 480). A system is simple to the extent that it comprises fewer axioms, or those axioms have simple forms (e.g. linear equations are simpler than polynomials of degree greater than 1). Often greater strength can be achieved at a cost in terms of simplicity (e.g. by adding axioms), and vice versa.

A candidate system may sometimes achieve a good deal of strength with little cost in simplicity if it is endowed with a probability function (cp. Loewer 2004, p. 1,119): that is, a function $Ch_t(p)$ that maps propositions and time pairs $\langle p, t\rangle$ onto real values in the [0, 1] interval, and that obeys the axioms of probability.[1] This is where Lewis's third theoretical desideratum comes it: a system *fits* the actual course of history well to the extent that the associated probability function assigns a high probability to the actual course of history: the higher the probability, the better the fit (Lewis 1994, p. 480).[2]

[1] Or alternatively a Renyi-Popper measure $Ch(p|q)$ that maps proposition pairs $\langle p, q\rangle$ onto the reals in the [0, 1] interval. (Plausibly, if one takes conditional chance as basic in this way, then it is redundant to include a "time" index to the chance function; see Hoefer 2007, pp. 562–565; Glynn 2010, pp. 78–79.)

[2] This notion of fit applies only if there are finitely many chance events. See Elga (2004) for an extension to infinite cases. In addition, if one wants to allow the possibility of statistical mechanical probabilities counting as chances, then one needs a notion of fit according to which one way a system may fit better is if its probability function assigns a relatively high probability to the macro-history of the world conditional upon a coarse-graining of its initial conditions (as well as assigning a relatively high probability to the micro-history conditional upon a fine graining of the initial conditions).

The best system is that which strikes the best balance between the theoretical virtues of simplicity, strength, and fit. According to the BSA, the probability function associated with the best system is the *chance function* for the world. The laws are the (axioms and) theorems of the best system.

The idea, then, is that the Humean mosaic, together with the theoretical virtues, serves to fix a best system. As Lewis (*ibid.*) puts it: "The arrangement of qualities provides the candidate ... systems, and considerations of simplicity and strength [and fit] and balance do the rest". Or, more concisely, chances are "Humean Best System-supervenient on the Humean mosaic" (Frigg and Hoefer 2013).

Lewis himself acknowledges that the BSA is not completely unproblematic: "[t]he worst problem about the best-system analysis" (Lewis 1994, p. 479) is that notions such as simplicity, strength, and balance are to some extent *imprecise*. There is, for example, no unique and maximally determinate simplicity metric that is obviously the correct one to apply to candidate systems, nor is there a unique and maximally determinate correct exchange rate between the competing virtues of simplicity, strength, and fit. The worry is that, within acceptable ranges, different precisifications of the simplicity metric and of the exchange rate between the virtues will yield different verdicts about which system counts as *best*. Lewis has little more to offer than the hope that this will not turn out to be so:

> If nature is kind, the best system will be *robustly* best–so far ahead of its rivals that it will come out first under any standards of simplicity and strength and balance. We have no guarantee that nature is kind in this way, but no evidence that it isn't. It's a reasonable hope. Perhaps we presuppose it in our thinking about law. I can admit that *if* nature were unkind, and *if* disagreeing rival systems were running neck-and-neck, then ... the theorems of the barely-best system would not very well deserve the name of laws. But I'd blame the trouble on unkind nature, not on the analysis; and I suggest we not cross these bridges unless we come to them. (Lewis *op cit*., p. 479; italics original)

One *might* think that the same thing that Lewis says about laws should be said of chances: if there is no clear winner of the best system competition, then there would be nothing deserving of the name chance. (Lewis himself, however, does not explicitly say this.) As we shall see in Sect. 5, Beisbart's central thesis is that we *do* in fact have good reason to think that there is a set of rival systems – each of which is associated with a different probability function – that are running neck-and-neck in the best system competition for our world. Beisbart draws the conclusion that, for the Best System analyst, there simply is nothing deserving of the name of chance.

3 The BSA and Statistical Mechanics

Lewis himself appears to have thought that the probability function associated with the best system for our world would simply be the quantum mechanical probability function: that is, the function that yields all and only the probabilities entailed by quantum mechanics, or whatever fundamental physical theory replaces it (see Lewis 1986, p. 118; 1994).

Yet Loewer (2001, 2007, 2008, 2012a, b) has influentially argued that the probabilities of statistical mechanics (SM) can also be understood as probabilities of the best system, and therefore as genuine objective chances on the BSA. Loewer appeals to the axiomatization of SM described by Albert (2000, chs. 3–4). Albert suggests that SM can be derived from the following:

(FD) the fundamental dynamical laws
(PH) a proposition characterizing the initial conditions of the universe as constituting a special low-entropy state; and
(SP) a uniform probability distribution (on the standard Lebesgue measure) over the regions of microphysical phase space associated with that low-entropy state.[3]

Albert (2012) and Loewer (2012a, b) dub the conjunction FD & PH & SP "the Mentaculus".

The argument that the SM probabilities are derivable from the Mentaculus goes roughly as follows. Consider the region of microphysical phase space associated with the low-entropy initial state of the universe implied by PH. Relative to the total volume of that region, the volume taken up by microstates that lead (by FD) to fairly sustained entropy increase until thermodynamic equilibrium is reached (and to the universe staying at or close to equilibrium thereafter) is extremely high. Consequently, the uniform probability distribution (given by SP) over the entire region yields an extremely high probability of the universe following such a path. When it comes to (approximately) isolated subsystems of the universe the idea is that, since a system's becoming approximately isolated is not itself correlated with its initial microstate being entropy-decreasing, it is extremely likely that any such subsystem that is in initial disequilibrium will increase in entropy over time (see Loewer 2007, p. 302, 2012a, pp. 124–125; 2012b, p. 17; Albert 2000, pp. 81–85).[4]

Albert (2000, 2012) and Loewer (2007, 2008, 2012a, b) have argued that the Mentaculus entails many of the probabilities of the special sciences. In virtue of this, Loewer claims that the Mentaculus is much stronger than a system consisting of the fundamental dynamical laws, FD, alone (Loewer 2012a, p. 129 and Frisch (forthcoming)). And since it is not much more complicated (it only requires the addition of the axioms PH and SP), Loewer claims that it is a plausible *best* system for our world (*ibid.*; also Loewer 2001, p. 618).[5]

[3] In the quantum case, the uniform probability distribution is not over classical phase space, but over the set of quantum states compatible with the PH.

[4] Though see Winsberg (2004) and Earman (2006) for criticisms of this line or argument.

[5] This proposal requires that initial conditions, such as PH, are potential axioms of the best system. The BSA has not always been construed as allowing for this. However, Lewis (1983, p. 367) himself seems sympathetic to the view that they may be.

4 Beisbart's Response

Let us assume that Loewer is correct that the Mentaculus constitutes a better system than one comprising the fundamental dynamics alone, and that it entails the SM probabilities. If the Mentaculus comes out *best*, then the SM probabilities will count as objective chances on the BSA.

But an axiom system consisting of *only* the fundamental dynamic laws is not the only rival to the Mentaculus. Schaffer (2007, pp. 130–132), Hoefer (2007, p. 560), and Beisbart (2014) consider another candidate, which consists of the fundamental dynamic laws plus an axiom giving the *precise initial conditions* of the universe. This is a very strong system. Schaffer (2007, pp. 131–132) suggests that it is maximally strong, while Hoefer (2007, p. 560) questions this. Hoefer (*ibid.*) points out that it's not obvious how to quantify the complexity of the two candidate systems in such a way as to allow comparison. It is also not obvious how to decide whether any increase in complexity is worth it because of the strength thereby bought (see Frisch 2011 and Frisch (forthcoming)).

Beisbart (2014) suggests that there are still further competitors to the Mentaculus.[6] As Beisbart observes:

> We can improve fit when we ... assume a flat probability distribution over a certain sub-region of the past low-entropy macro-state [as opposed to over the whole of the past low-entropy macro-state, as per (SP) of the Mentaculus]. That sub-region may be defined by the demand that a certain elementary particle has a kinetic energy larger than a particular value e_0, for instance.
>
> If we do so, we have to pay in simplicity though because [in addition to the assumed low-entropy initial state, we have to further specify that the initial] kinetic energy of a particular particle be e_0.
>
> So overall, would we improve the system ... ? This is a difficult question, and the answer is far from clear. The only thing we can say is that fit is considerably improved, but that there is a considerable cost in simplicity too. So it's not a case in which the right sort of balance favors one system rather than the other in a clear way.

Beisbart's worry is that, depending upon which sub-region of the phase space associated with PH the flat distribution is applied to, we get a range of candidate best systems (cp. also Schaffer 2007, p. 131n). At the one extreme, we have the

[6] Callender (2011) calls attempts to derive the SM probabilities from a probability distribution over the initial conditions of *the universe as a whole* "Globalist" approaches to axiomatizing SM. The Mentaculus is one example of a Globalist approach. Beisbart points out that there are rivals. In contrast to Globalist approaches, "Localist" approaches (see Callender *op cit.*) attempt to derive the SM probabilities from probability distributions over the initial states of the various *approximately isolated subsystems of the universe*. Beisbart observes that there is a range of competing Localist approaches to axiomatizing SM. While, for reasons of space, we will here focus upon the competing Globalist approaches, many of our points will carry across to the competition between Localist axiomatizations, if one thinks that the Localist approaches are more promising (see Glynn unpublished).

Mentaculus (where a uniform distribution is applied to the whole region of phase space compatible with PH)[7]; at the other extreme, we have a system comprising the fundamental dynamic laws together with the precise initial conditions. The latter is equivalently to what we get in the limit as we apply a uniform distribution to smaller and smaller sub-regions of the phase space associated with PH, each of which contains the point-sized region of phase space that the universe actually initially occupied. The former is very simple, but gives an inferior fit; the latter gives a better fit, but is less simple. In between we have a continuum of systems involving the application of a uniform distribution to progressively smaller sub-regions of the phase-space compatible with PH (where each sub-region contains the actual point in phase space at which our universe was initially located). Such systems are increasingly better fitting, since they assign an increasingly high probability to the actual macroscopic course of events, but also increasingly complex, since picking out progressively smaller sub-regions requires building into the axioms an increasing amount of information about the actual initial state of the universe.

If we had a precise simplicity measure, and a precise exchange-rate between simplicity and fit, then perhaps the measure and the exchange rate could produce an exact tie between systems located on this continuum. The idea would be that, according to the exchange rate, the change in fit as we move along the continuum is precisely counterbalanced by the change in simplicity. More plausibly, a precise simplicity measure and a precise exchange rate is something we cannot reasonably hope to have. If so, it is quite plausible that none of the systems on this continuum is robustly better than all – or indeed *any* – of the others. That is, none is superior to all – or *any* – others given the imprecision of simplicity and of the exchange rate. It is on these grounds that Beisbart argues that:

> The conclusion looming large here is that Humean considerations do not provide us with any clearly optimal solution to the problem of how to define a dynamics of chances. And good just isn't good enough. We need a best system that is clearly best, and not just a good one. If there isn't a best system, then there are no chances ....

In other words, Beisbart's idea is that the Humean mosaic, together with the theoretical virtues, fails to single out a unique best system, and therefore a corresponding probability function. Consequently he claims that the BSA implies that there is nothing that counts as the *objective chance function*. This is precisely analogous to Lewis's claim that there would be nothing deserving of the name of *law* if several systems were roughly tied for *best*.

[7] Indeed, as Beisbart points out, it is not clear precisely how *low* initial entropy is specified to be by the PH. Different precisifications of the PH yield different sized regions of phase space to which the uniform distribution is to be applied. So it seems that the number of competing systems may be larger still.

5 Humean Chances Aren't Sharp

Yet Beisbart's ultimate conclusion can be resisted. Firstly, one might wonder whether it is really the case that the choice of initial phase space region (and of initial probability distribution) significantly affects the probabilities that systems exhibit thermodynamic-like behavior. While there may not be a unique best system, the range of best systems might nevertheless all agree on the probabilities that they assign to macro-events.[8] This is a possibility that Beisbart himself notes (pp. 16 & 18 of draft). Yet even if different systems roughly tied for first place assign different probabilities to the macroscopic course of events, we do not agree that this implies that there can be nothing deserving of the name *chance*.[9] The fact that the Humean mosaic, together with the (imprecise) relation of Humean Best System supervenience does not uniquely fix a single axiom system should not lead the Humean to deny that there are chances in the world. Rather, it should lead her to deny that the relevant chances must be sharp. If there is no clear winner in the best system competition, it seems to us that the natural thing for the Humean to say is that the *set* of probability functions corresponding to the tied systems constitute the set of chance functions for the world. Where the probability functions for the tied systems agree on the probability for a particular event then the objective chance for that event is sharp. This seems quite possible when, for example, we are considering micro-physical events like the decay of a tritium atom within the next 12.32 years. On the other hand, when the probability functions of the tied-for-best systems yield a range of values then the range constitutes an unsharp chance for the event in question.

The basic intuition behind the introduction of unsharp Humean chances in this context is that it seems unreasonable in the case of a tie between systems that there simply is *nothing* playing the correct role in guiding rational credence, and thus nothing serving as a "guide to life" for dealing with the relevant situations. Hoefer (2007, pp. 580–587) and Frigg and Hoefer (2010) argue that Humean chances are constraints on rational credence because of the tight connection between Humean chances and actual frequencies. Specifically they argue that this allows for a "consequentialist" justification of the chance-credence connection: given the tight connection between Humean chances and actual frequencies, agents betting

[8] At least this might be so if, with Frigg and Hoefer (2013), we exclude information about the precise micro-state of the world as inadmissable, since "*chance rules operate at a specific level and evidence pertaining to more fundamental levels is inadmissible.*" See Maudlin (2007) for a discussion of how one can derive typical thermodynamic behavior without committing to precise assumptions either about the initial probability distribution or the size of the initial phase space region.

[9] As we saw, Lewis claims that if there was not a unique best system, there would be nothing deserving of the name *law*, though he earlier (Lewis 1983, p. 367) said theorems entailed by *all* of the tied systems would count as laws. We're sympathetic to his earlier position. If, for instance, the fundamental dynamics are the same in all the tied systems (this is not something that is disputed by Beisbart and others who have examined rivals to the Mentaculus), then they will come out as laws.

according to the Humean chances will do well in the long run. The assumption that there is a unique probability function that serves as the Humean chance function does not seem essential to this argument.[10] In the case of a tie between systems, it seems that one can analogously argue that the set of probability functions entailed by the tied systems would also constrain reasonable credence. Specifically, when confronted with a tie between systems, an agent who knew the *set* of probability functions corresponding to the tied systems, and who knew that a certain chance setup is instantiated, and who had no inadmissible information (no information beyond knowledge of the initial chance setup and the set of probability distributions entailed by the tied-for-best systems), rationally ought *not* to adopt a credence value that lies outside the range of probability values entailed by the systems that are tied for *best*.

For example, any tied-for-best system will entail a very low probability for entropy decrease in isolated systems. A system that does not would be straightforwardly excluded from the set of tied-for-best systems: it will inevitably be highly ill-fitting, since fit implies a close correspondence to the actual frequencies. We can thus offer a "consequentialist" justification for not adopting a credence outside of the set of probabilities yielded by the tied systems: betting as though entropy-decrease is not very improbable would lead one to do very badly in the long run.

Indeed, rather plausibly, in such a case, a reasonable agent would have a credence assignment which was unsharp: specifically, it would be represented by a set of values corresponding to those entailed by the probability functions of the tied systems. In such situations an agent would have no rational basis to choose between the probability functions entailed by the tied systems but would be rationality compelled to base their behaviour upon the relevant unsharp chance. Thus, in such a situation the set of values entailed by the tied systems is playing the key chance-role of guiding reasonable credence, and thereby constitutes an unsharp Humean chance.[11]

There is one worry here. Given a set of systems that are tied for best – because some have greater fit, while others have greater simplicity – it would appear to be (instrumentally) rational to set one's credences according to the probabilities entailed by that system which has the greatest fit (i.e. accords best with the frequencies), since betting according to those probabilities would yield the greatest payoffs in the long run. We agree that this is a worry, but it is a general problem for Best System analyses of chance. Even in the case of a unique best system, the best system chances are liable to depart somewhat from the actual frequencies. This is because both simplicity and fit go into determining a best system. But it seems difficult to argue that the best system probabilities, rather than the actual frequencies, are the best players of the chance role in guiding rational credence: if one knew both,

[10]Frigg and Hoefer (2013) themselves find it plausible that there may not be a unique best system for our world.

[11]Elga (2010) argues that unsharp credences are incompatible with perfect rationality. If this were correct, then perhaps any player of the chance role in guiding rational credence must itself be sharp. However, we find Joyce's (2010) defense of unsharp credences against Elga's argument compelling.

one would do better in the long run if one bet according to the actual frequencies. The best system analysis of chance requires some general explanation of why the best-fitting probabilities aren't automatically the chances (perhaps, for example, one could appeal to other aspects of the chance role). Whatever answer is deployed in this context will also be available to us in explaining why the best fitting of the tied systems doesn't automatically deliver the chances and so, while we admit this issue to be problematic for our approach, we do not take it to be unduly or uniquely so (see Glynn unpublished).

Let us now go back and reconsider the tied-for-best systems discussed in the previous section in the context of unsharp probabilities. Consider again the Mentaculus, i.e. FD & PH & SP. Given FD, as defined above, there are two things that need to be fixed: the initial macrostate that the universe is in, and the probability distribution over the associated region of microphysical phase space. The Mentaculus requires the macrostate to be a special low-entropy macrostate, M_0, which is associated with a phase space region, Γ_0. It requires the probability distribution over Γ_0 to be the uniform distribution ρ_U. The discussed tied-for-best systems vary either the size of the initial low-entropy region or the probability distribution over that region. In many cases, systems that vary the region will be equivalent to systems that vary the probability distribution. So, for now, we consider only changes to the size of the initial low-entropy region. Beisbart, as discussed, considers a sub-region Γ_B of the phase space region Γ_0 associated with the low entropy macrostate M_0 specified by the Mentaculus. This sub-region Γ_B is defined by adding to the requirement that the region be associated with a low-entropy macrostate, the specification that one of the elementary particles comprising the initial universe has kinetic energy above a certain value. As discussed in the last section, one can consider a continuum of sub-regions of Γ_0, given by specifying more and more of the microphysical details of the universe's initial state. As we move along this continuum, we get better and better fit at the cost of less and less simplicity up to the extreme case where we specify the precise initial condition, given by the region Γ_δ, which contains only one point: namely the exact microphysical state of the universe.

We can now formulate a family of Mentaculus-like axiomatizations of SM, which we call M(Γ). Members of the family M(Γ) comprise FD plus the following:

(PH) a proposition characterizing the initial condition of the universe as some particular element of the set $m = \{M_0, \ldots M_B, \ldots M_\delta\}$; and

(SP) a probability distribution $\rho = \rho_U$ (on the standard Lebesgue measure) over the associated region $\gamma = \{\Gamma_0, \ldots, \Gamma_B, \ldots\}$ of microphysical phase space.

(SP is redundant for the case where the precise initial condition is taken.) The suggestion is that SM probabilities are entailed by M(Γ) for each choice of $\Gamma \in \gamma$, and that, by considering all $\Gamma \in \gamma$, we obtain SM probabilities that are set-valued, i.e. are unsharp.

There are several issues that need to be addressed. First, one can obviously generalize the above to include a variety of probability distributions, i.e. to consider M(Γ, ρ), if axiom systems containing different ρs are among those that are

tied-for-best. Whether such a generalization is necessary, given the existence of the family M(Γ) of tied-for-best systems, is questionable. Consider, for example, Frigg and Hoefer's (2013) suggestion of "a peaked distribution, nearly Dirac-delta style" over the precise initial condition represented by ρ_δ. In effect, $M(\Gamma_0,\rho_\delta) = M(\Gamma_\delta)$. Second, we have only considered subregions of the low-entropy macro-state that are in γ. One could also consider regions larger than Γ_0 and may find systems that are tied for best.

But is entropy then still low enough to provide a least a minimally good fit with the actual course of history? How low is low enough is a non-trivial question and one seems to be on the safe side to consider only sub-regions of Γ_0, which by construction is low enough.[12] Third, the exact relation between M(Γ) and the SM probabilities is a difficult question. Whether, for instance, a continuum of subregions translates into a continuum of probabilities for SM is far from trivial and is a question for detailed calculations.

While each of these three issues undoubtedly warrant further detailed attention, we feel that they are indicative of the potential for fruitful refinement of the idea of unsharp Humean chances. Some of these issues are taken up further in Glynn (unpublished).

6 Conclusion

Beisbart's paper is a useful addition to the literature on Statistical Mechanics and the BSA. In particular, it bears emphasis that the notions of simplicity, strength, and balance are imprecise and that plausibly, in conjunction with the Humean mosaic, do not single out a single best system, but rather a set of tied-for-best systems. But we don't think that, from the existence of such tied-for-best systems, one is justified in concluding the non-existence of Humean chances. Rather, such cases can plausibly be construed as implying the existence of *non-sharp Humean chances*. In this brief reply, we have offered an outline of such a proposal. Detailed development of this view must await another occasion.

[12] Also it's not clear to us that an axiom system specifying that the universe was initially in a larger region of phase space than Γ_0 is simpler than one specifying that the universe was in Γ_0. As is well known (e.g. Lewis 1983, p. 367), simplicity is vocabulary-relative and, in order to avoid trivializing the desideratum of simplicity, we must take simplicity-when-formulated-with-unnatural-predicates to be less desirable than simplicity-when-formulated-with-reasonably-natural-predicates (and perhaps simplicity-when-formulated-with-perfectly-natural-predicates to be more desirable than both). In our not-too-unnatural macro-vocabulary, we may be able to formulate simple axioms that pick out moderately large regions of phase space like Γ_0. Picking out smaller regions will often require employing complex microphysical predicates, thus increasing fit (and perhaps naturalness of predicates) at the expense of simplicity. But picking out larger regions than Γ_0 may require disjunctions of reasonably natural macrophysical predicates. The resulting axiom systems will be both worse fitting and less simple (or worse fitting and formulated in less natural language). If so, they would be clearly inferior to an axiom system specifying that the universe was initially in Γ_0.

Acknowledgement The authors would like to thank Claus Beisbart, Seamus Bradley and Leszek Wroński for helpful comments. We would also like to acknowledge the support of the Alexander von Humboldt-Foundation and the Munich Center for Mathematical Philosophy.

References

Albert, D. 2000. *Time and chance*. Cambridge, MA: Harvard University Press.

Albert, D. 2012. Physics and chance. In *Probability in physics*, ed. Y. Ben-Menahem and M. Hemmo, 17–40. Berlin: Springer.

Beisbart, C. 2014. Good just isn't good enough – Humean chances and Boltzmannian statistical physics. In *New directions in the philosophy of science*, ed. M.C. Galavotti and D. Dieks. Dordrecht: Springer.

Callender, C. 2011. The past histories of molecules. In *Probabilities in physics*, ed. C. Beisbart and S. Hartmann, 83–113. New York: Oxford University Press.

Earman, J. 2006. The 'past hypothesis': Not even false. *Studies in History and Philosophy of Modern Physics* 37: 399–430.

Elga, A. 2004. Infinitesimal chances and the laws of nature. *Australasian Journal of Philosophy* 82: 67–76.

Elga, A. 2010. Subjective probabilities should be sharp. *Philosopher' Imprint* 10: 1–11.

Fenton-Glynn, L. unpublished. Unsharp best system chances. http://philsci-archive.pitt.edu/10239/

Frigg, R., and C. Hoefer. 2013. The best humean system for statistical mechanics. *Erkenntnis*. doi:10.1007/s10670-013-9541-5.

Frigg, R., and C. Hoefer. 2010. Determinism and chance from a humean perspective. In *The present situation in the philosophy of science*, ed. D. Dieks, W.J. Gonzalez, S. Hartmann, M. Weber, F. Stadler, and T. Uebel, 351–371. Berlin/New York: Springer.

Frisch, M. 2011. From Boltzmann to Arbuthnot: Higher-level laws and the best system. *Philosophy of Science* 78: 1001–1011.

Frisch, M. forthcoming. Physical fundamentalism in a Lewisian best system. In *Asymmetries of chance and time*, ed. Alastair Wilson. Oxford University Press.

Glynn, L. 2010. Deterministic chance. *The British Journal for the Philosophy of Science* 61: 51–80.

Hoefer, C. 2007. The third way on objective chance: A sceptic's guide to objective chance. *Mind* 116: 549–596.

Joyce, J. 2010. A defense of imprecise credences in inference and decision making. *Philosophical Perspectives* 24: 281–323.

Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy* 61: 343–377.

Lewis, D. 1986. Postscripts to 'a subjectivist's guide to objective chance'. In his *Philosophical papers*, vol. 2, 114–132. Oxford: Oxford University Press.

Lewis, D. 1994. Humean supervenience debugged. *Mind* 103: 473–490.

Loewer, B. 2001. Determinism and chance. *Studies in History and Philosophy of Science* 32: 609–620.

Loewer, B. 2004. David Lewis's humean theory of objective chance. *Philosophy of Science* 71: 1115–1125.

Loewer, B. 2007. Counterfactuals and the second law. In *Causation, physics, and the constitution of reality: Russell's republic revisited*, ed. H. Price and R. Corry, 293–326. Oxford: Clarendon.

Loewer, B. 2008. Why there is anything except physics. In *Being reduced: New essays on reduction, explanation and causation*, ed. J. Hohwy and J. Kallestrup, 149–163. Oxford: Oxford University Press.

Loewer, B. 2012a. Two accounts of laws and time. *Philosophical Studies* 160: 115–137.

Loewer, B. 2012b. The emergence of time's arrows and special science laws from physics. *Interface Focus* 2: 13–19.

Maudlin, T. 2007. What could be objective about probabilities? *Studies in History and Philosophy of Modern Physics* 38: 275–291.

Schaffer, J. 2007. Deterministic chance? *The British Journal for the Philosophy of Science* 58: 113–140.

Winsberg, E. 2004. Can conditioning on the 'past hypothesis' militate against the reversibility objections? *Philosophy of Science* 71: 489–504.

Noncommutative Causality in Algebraic Quantum Field Theory

Gábor Hofer-Szabó

1 Introduction

Algebraic quantum field theory (AQFT) is a mathematically transparent quantum theory with clear conceptions of locality and causality (see Haag 1992 and Halvorson 2007). In this theory observables are represented by a net of local C^*-algebras associated to bounded regions of a given spacetime. This correspondence is established due to the axioms of the theory such as isotony, microcausality and covariance. A state ϕ in this theory is defined as a normalized positive linear functional on the quasilocal observable algebra $\mathcal{A}$ which is the inductive limit of local observable algebras. The representation $\pi_\phi : \mathcal{A} \to \mathcal{B}(\mathcal{H})$ corresponding to the state ϕ transforms the net of C^*-algebras into a net of von Neumann observable algebras by closures in the weak topology.

In AQFT *events* are typically represented by projections of a von Neumann algebra. Although due to the axiom of microcausality two projections A and B commute if they are contained in local algebras supported in spacelike separated regions, they can still *be correlating* in a state ϕ, that is

$$\phi(AB) \neq \phi(A)\phi(B) \tag{1}$$

in general. In this case the correlation between these events is said to be *superluminal*. A remarkable characteristics of Poincaré covariant theories is that there exist "many" normal states establishing superluminal correlations (for the precise meaning of "many" see Summers and Werner 1988 and Halvorson and Clifton 2000). Since spacelike separation excludes direct causal influence, one may look for a causal explanation of these superluminal correlations in terms of *common causes*.

G. Hofer-Szabó (✉)
Research Centre for the Humanities, Institute of Philosophy, Úri u. 53, 1014 Budapest, Hungary
e-mail: szabo.gabor@btk.mta.hu

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_38,

The first probabilistic definition of the common cause is due to Hans Reichenbach (1956). Reichenbach characterizes the notion of the common cause in the following probabilistic way. Let (Σ, p) be a classical probability measure space and let A and B be two positively correlating events in Σ that is let

$$p(A \wedge B) > p(A)\, p(B). \tag{2}$$

Definition 1. An event $C \in \Sigma$ is said to be the *common cause* of the correlation (A, B) if the following conditions hold:

$$p(A \wedge B|C) = p(A|C)p(B|C) \tag{3}$$

$$p(A \wedge B|C^{\perp}) = p(A|C^{\perp})p(B|C^{\perp}) \tag{4}$$

$$p(A|C) > p(A|C^{\perp}) \tag{5}$$

$$p(B|C) > p(B|C^{\perp}) \tag{6}$$

where $C^{\perp}$ denotes the orthocomplement of C and $p(\cdot|\cdot)$ is the conditional probability.

The above definition, however, is too specific to be applied in AQFT since (i) it allows only for causes with a *positive* impact on their effects, (ii) it excludes the possibility of a *set* of cooperating common causes, (iii) it is silent about the spatiotemporal *localization* of the events and (iv) most importantly, it is *classical*. Therefore we need to generalize Reichenbach's original definition of the common cause. For the sake of brevity, we do not repeat here all the intermediate steps of the entire definitional process (for this see Hofer-Szabó and Vecsernyés 2012a), but jump directly to the most general definition of the common cause in AQFT.

Let $\mathcal{P}(\mathcal{N})$ be the non-distributive lattice of projections (events) in a von Neumann algebra $\mathcal{N}$ and let $\phi\colon \mathcal{N} \to \mathbb{C}$ be a state on it. A set of mutually orthogonal projections $\{C_k\}_{k\in K} \subset \mathcal{P}(\mathcal{N})$ is called a *partition of the unit* $\mathbf{1} \in \mathcal{N}$ if $\sum_k C_k = \mathbf{1}$. Such a partition defines a *conditional expectation*

$$E\colon \mathcal{N} \to \mathcal{C},\ A \mapsto E(A) := \sum_{k\in K} C_k A C_k, \tag{7}$$

that is a unit preserving positive surjection onto the unital C^*-subalgebra $\mathcal{C} \subseteq \mathcal{N}$ obeying the bimodule property $E(B_1 A B_2) = B_1 E(A) B_2; A \in \mathcal{N}, B_1, B_2 \in \mathcal{C}$. We note that $\mathcal{C}$ contains exactly those elements of $\mathcal{N}$ that commute with $C_k, k \in K$. Recall that $\phi \circ E$ is also a state on $\mathcal{N}$.

Now, let $A, B \in \mathcal{P}(\mathcal{N})$ be two commuting events correlating in state ϕ in the sense of (1). (We note that in case of projection lattices we will use only algebra operations (products, linear combinations) instead of lattice operations $(\vee, \wedge)$. In case of commuting projections $A, B \in \mathcal{P}(\mathcal{N})$ we have $A \wedge B = AB$ and $A \vee B = A + B - AB$.)

Definition 2. A partition of the unit $\{C_k\}_{k\in K} \subset \mathcal{P}(\mathcal{N})$ is said to be a *common cause system* of the correlation (1) if

$$\frac{(\phi\circ E)(ABC_k)}{\phi(C_k)} = \frac{(\phi\circ E)(AC_k)}{\phi(C_k)}\frac{(\phi\circ E)(BC_k)}{\phi(C_k)} \tag{8}$$

for $k \in K$ with $\phi(C_k) \neq 0$. If C_k commutes with both A and B for all $k \in K$ we call $\{C_k\}_{k\in K}$ a *commuting* common cause system, otherwise a *noncommuting* one. A common cause system of size $|K| = 2$ is called a *common cause*. Reichenbach's definition (without the inequalities (5) and (6)) is a commuting common cause in the sense of (8).

Some remarks are in place here. First, in case of a commuting common cause system $\phi\circ E$ can be replaced by ϕ in (8) since $(\phi\circ E)(ABC_k) = \phi(ABC_k), k \in K$. Second, using the decompositions of the unit, $\mathbf{1} = A + A^\perp = B + B^\perp$, (8) can be rewritten in an equivalent form:

$$(\phi\circ E)(ABC_k)(\phi\circ E)(A^\perp B^\perp C_k) = (\phi\circ E)(AB^\perp C_k)(\phi\circ E)(A^\perp BC_k),\ k \in K. \tag{9}$$

One can even allow here the case $\phi(C_k) = 0$ since then both sides of (9) are zero. Third, it is obvious from (9) that if $C_k \leq X$ with $X = A, A^\perp, B$ or $B^\perp$ for all $k \in K$, then $\{C_k\}_{k\in K}$ serves as a (commuting) common cause system of the given correlation independently of the chosen state ϕ. Hence, these solutions are called *trivial common cause systems*. If $|K| = 2$, triviality means that $\{C_k\} = \{A, A^\perp\}$ or $\{C_k\} = \{B, B^\perp\}$. Obviously, for superluminal correlation one looks for nontrival common causal explanations.

In AQFT one also has to specify the spacetime localization of the common causes. They have to be in the past of the correlating events. But in which past? One can define different pasts of the bounded regions V_A and V_B in a given spacetime as:

$$\begin{aligned} \textit{weak past:}\quad & wpast(V_A, V_B) := I_-(V_A) \cup I_-(V_B) \\ \textit{common past:}\quad & cpast(V_A, V_B) := I_-(V_A) \cap I_-(V_B) \\ \textit{strong past:}\quad & spast(V_A, V_B) := \cap_{x\in V_A\cup V_B} I_-(x) \end{aligned}$$

where $I_-(V)$ denotes the union of the backward light cones $I_-(x)$ of every point x in V (Rédei and Summers 2007). Clearly, $wpast \supset cpast \supset spast$.

With all these definitions in hand we can now define six different common cause systems in local quantum theories according to (i) whether *commutativity* is required and (ii) whether the common cause system is localized in the *weak*, *common* or *strong* past. Thus we can speak about *commuting/noncommuting (weak/strong) common cause systems*.

To address the EPR-Bell problem we will need one more concept. In the EPR scenario the real challenge is to provide a common causal explanation *not* for

one *single* correlating pair but for a *set* of correlations (typically three or four correlations). Therefore, we also need to introduce the notion of the so-called joint[1] common cause system:

Definition 3. Let $\{A_m; m = 1, \ldots M\}$ and $\{B_n; n = 1, \ldots N\}$ be finite sets of projections in the algebras $\mathcal{A}(V_A)$ and $\mathcal{A}(V_B)$, respectively, supported in spacelike separated regions V_A and V_B. Suppose that all pair of spacelike separated projections (A_m, B_n) correlate in a state ϕ of $\mathcal{A}$ in the sense of (1). Then the set $\{(A_m, B_n); m = 1, \ldots M; n = 1, \ldots N\}$ of correlations is said to possess a commuting/noncommuting (weak/strong) *joint* common cause system if there exists *a single* commuting/noncommuting (weak/strong) common cause system for *all* correlations (A_m, B_n).

Since providing a *joint* common cause system for a set of correlations is much more demanding than simply providing a common cause system for a *single* correlation, therefore we keep the question of the common causal explanation separated from that of the *joint* common causal explanation. In Sect. 2 we will investigate the possibility of a common causal explanation for a *single* correlation—or in the philosophers' jargon, the status of Reichenbach's famous Common Cause Principle in AQFT. In Sect. 3 we will address the more intricate question as to whether EPR correlations can be given a *joint* common causal explanation. The crucial common element in both sections will be *noncommutativity*. We will argue that embracing *noncommuting* common causes in our causal explanation helps us in both cases: (i) in the case of common causal explanation it helps to maintain the validity of Reichenbach's Common Causal Principle in AQFT; (ii) in the case of *joint* common causal explanation it helps to provide a local, *joint* common causal explanation for a set of correlations violating the Bell inequalities. We conclude the paper in Sect. 4.

2 Noncommutative Common Cause Principles in AQFT

Reichenbach's Common Cause Principle (CCP) is the following metaphysical claim: if there is a correlation between two events and there is no direct causal (or logical) connection between the correlating events, then there exists a common cause of the correlation. The precise definition of this informal statement that fits to AQFT is the following:

Definition 4. A local quantum theory is said to satisfy the Commutative/ Noncommutative (Weak/Strong) CCP if for any pair $A \in \mathcal{A}(V_A)$ and $B \in \mathcal{A}(V_B)$ of projections supported in spacelike separated regions V_A, V_B and for every locally faithful state $\phi\colon \mathcal{A} \to \mathbb{C}$ establishing a correlation between A and B in the sense

[1] In Hofer-Szabó and Vecsernyés (2012a, 2013a) called *common* common cause system.

of (1), there exists a *nontrivial* commuting/noncommuting common cause system $\{C_k\}_{k\in K} \subset \mathcal{A}(V)$ such that the localization region V is in the (weak/strong) common past of V_A and V_B.

What is the status of these six different CCPs in AQFT?

The question as to whether the Commutative CCPs are valid in a Poincaré covariant local quantum theory in the von Neumann algebraic setting was first raised by Rédei (1997, 1998). As a positive answer to this question, Rédei and Summers (2002, 2007) have shown that the Commutative Weak CCP holds in algebraic quantum field theory with locally infinite degrees of freedom in the following sense: for every locally normal and faithful state and for every superluminally correlating pair of projections there exists a weak common cause, that is a common cause system of size 2 in the weak past of the correlating projections. They have also shown that the localization of a common cause cannot be restricted to $wpast(V_A, V_B) \setminus I_-(V_A)$ or $wpast(V_A, V_B) \setminus I_-(V_B)$ due to logical independence of spacelike separated algebras.

Concerning the Commutative (Strong) CCP less is known. If one also admits projections localized only in *un*bounded regions, then the Strong CCP is known to be false: von Neumann algebras pertaining to complementary wedges contain correlated projections but the strong past of such wedges is empty (see Summers and Werner 1988 and Summers 1990). In spacetimes having horizons, e.g. those with Robertson–Walker metric, there exist states which provide correlations among local algebras corresponding to spacelike separated bounded regions such that the common past of these regions is again empty (Wald 1992). Hence, CCP is not valid there. Restricting ourselves to *local* algebras in Minkowski spaces the situation is not clear. We are of the opinion that one cannot decide on the validity of the (Strong) CCP without an explicit reference to the dynamics.

Coming back to the proof of Rédei and Summers, the proof had a crucial premise, namely that the algebras in question are *von Neumann algebras of type III*. Although these algebras are the typical building blocks of Poincaré covariant theories, other local quantum theories apply von Neumann algebras of other type. For example, theories with locally finite degrees of freedom are based on von Neumann algebras of type I. This raised the question as to whether the Commutative Weak CCP is generally valid in AQFT. To address the problem Hofer-Szabó and Vecsernyés (2012a) have chosen a specific local quantum field theory, the local quantum Ising model having locally finite degrees of freedom. It turned out that the Commutative Weak CCP does *not* hold in the local quantum Ising model and it cannot hold either in theories with locally finite degrees of freedom in general.

But why should we require commutativity between the common cause and its effects at all?

Commutativity has a well-defined role in any quantum theories. In standard quantum mechanics observables should commute to be simultaneously measurable. In AQFT the axiom of microcausality ensures that observables with spacelike separated supports—roughly, events happening "simultaneously"—commute. But cause and effect are typically *not* such simultaneous events! If one considers

ordinary QM, one well sees that observables do not commute even with their own time translates in general. For example, the time translate $x(t) := U(t)^{-1} x U(t)$ of the position operator x of the harmonic oscillator in QM does *not* commute with $x \equiv x(0)$ for generic t, since in the ground state vector ψ_0 we have

$$\big[x, x(t)\big]\,\psi_0 = \frac{-i\hbar \sin(\hbar\omega t)}{m\omega}\psi_0 \not\equiv 0. \tag{10}$$

Thus, if an observable A is not a conserved quantity, then the commutator $[A, A(t)] \neq 0$ in general. So why should the commutators $[A, C]$ and $[B, C]$ vanish for the events A, B and for their common cause C supported in their (weak/common/strong) past? We think that commuting common causes are only unnecessary reminiscence of their classical formulation. Due to their relative spacetime localization, that is due to the time delay between the correlating events and the common cause, it is also an unreasonable assumption.

Abandoning commutativity in the definition of the common cause is therefore a desirable move. The first benefit of allowing noncommuting common causes is that the noncommutative version of the result of Rédei and Summers can be regained. This result has been formulated in Hofer-Szabó and Vecsernyés (2013a) in the following:

Proposition 1. *The* Noncommutative *Weak CCP holds in local UHF-type quantum theories. Namely, if $A \in \mathcal{A}(V_A)$ and $B \in \mathcal{A}(V_B)$ are projections with spacelike separated supports V_A and V_B correlating in a locally faithful state ϕ on $\mathcal{A}$, then there exists a common cause $\{C, C^{\perp}\}$ localized in the weak past of V_A and V_B.*

Now, let us turn to the more complicated question as to whether a *set* of correlations violating the Bell inequality can have a *joint* common causal explanation in AQFT. Since our answer requires some knowledge of the main concepts of the Bell scenario in AQFT and some acquaintance with the model in which our results were formulated, we start the next section with a short tutorial on these issues (for more details see Hofer-Szabó and Vecsernyés 2012b, 2013b).

3 Noncommutative Joint Common Causal Explanation for Correlations Violating the Bell Inequality

The Bell problem is treated in AQFT in a subtle mathematical way (Summers and Werner 1987a,b; Summers 1990); here we introduce, however, only those concepts which are related to the problem of common causal explanation (for more on that see Hofer-Szabó and Vecsernyés 2013b).

Let $A_1, A_2 \in \mathcal{A}(V_A)$ and $B_1, B_2 \in \mathcal{A}(V_B)$ be projections with spacelike separated supports V_A and V_B, respectively. We say that in a locally faithful state

ϕ the Clauser–Horne-type *Bell inequality is satisfied* for A_1, A_2, B_1 and B_2 if the following inequality holds:

$$-1 \leqslant \phi(A_1B_1 + A_1B_2 + A_2B_1 - A_2B_2 - A_1 - B_1) \leqslant 0 \tag{11}$$

otherwise we say that the *Bell inequality is violated*. (Sometimes in the EPR-Bell literature another inequality, the so-called Clauser–Horne–Shimony–Holte-type Bell inequality is used as a constraint on the expectation of (not *projections* but) self-adjoint *contractions*. Since these two inequalities are equivalent, in what follows we will simply use (11) as *the* definition of the Bell inequality.)

In the literature it is a received view that if a set of correlations violates the Bell inequality, then the set cannot be given a joint common causal explanation. The following proposition proven in Hofer-Szabó and Vecsernyés (2013b) shows that this view is correct *only if* joint common causal explanation is meant as a *commutative* joint common causal explanation:

Proposition 2. *Let $A_1, A_2 \in \mathcal{A}(V_A)$ and $B_1, B_2 \in \mathcal{A}(V_B)$ be four projections localized in spacelike separated spacetime regions V_A and V_B, respectively, which correlate in the locally faithful state ϕ. Suppose that $\{(A_m, B_n); m, n = 1, 2\}$ has a joint common causal explanation in the sense of Definition 3. Then the following Bell inequality*

$$-1 \leqslant (\phi \circ E_c)(A_1B_1 + A_1B_2 + A_2B_1 - A_2B_2 - A_1 - B_1) \leqslant 0. \tag{12}$$

holds for the state $\phi \circ E_c$. If the joint common cause is a commuting *one, then the original Bell inequality (11) holds for the original state ϕ.*

Proposition 2 states that in order to yield a *commuting* joint common causal explanation for the set $\{(A_m, B_n); m, n = 1, 2\}$ the Bell inequality (11) has to be satisfied. This result is in complete agreement with the usual approaches to Bell inequalities (see e.g. Butterfield 1989, 1995, 2007). But what is the situation with *noncommuting* common cause systems? Since—apart from (12)—Proposition 2 is silent about the relation between a *noncommuting* joint common causal explanation and the Bell inequality (11), the question arises: can a *set* of correlations violating the Bell inequality (11) have a *noncommuting* joint common causal explanation?

In Hofer-Szabó and Vecsernyés (2012b, 2013b) it has been shown that the answer to the above question is positive: the violation of the Bell inequality does *not* exclude a joint common causal explanation *if* common causes can be noncommuting. Moreover, these common causes turned out to be localizable just in the "right" spacetime region (see below). For this result, we applied a simple AQFT with locally finite degrees of freedom, the so-called local quantum Ising model (for more details see Hofer-Szabó and Vecsernyés (2012b, 2013b); for a Hopf algebraic introduction of the model see Szlachányi and Vecsernyés (1993), Nill and Szlachányi (1997), Müller and Vecsernyés).

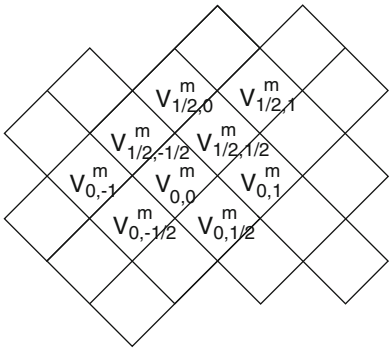

Fig. 1 The two dimensional discrete Minkowski spacetime covered by minimal double cones

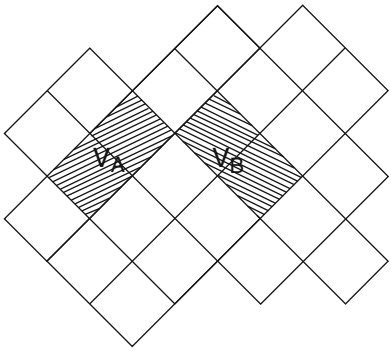

Fig. 2 Correlations between events in V_A and V_B

Consider a "discretized" version of the two dimensional Minkowski spacetime $\mathcal{M}^2$ covered by minimal double cones $V^m_{t,i}$ of unit diameter with their center in (t,i) for $t,i \in \mathbb{Z}$ or $t,i \in \mathbb{Z}+1/2$ (see Fig. 1). A non-minimal double cone $V_{t,i;s,j}$ in this covering can be generated by two minimal double cones in the sense that $V_{t,i;s,j}$ is the smallest double cone containing both $V^m_{t,i}$ and $V^m_{s,j}$. The set of double cones forms a directed poset which is left invariant by integer space and time translations.

The "one-point" observable algebras associated to the minimal double cones $V^m_{t,i}$ are defined to be $\mathcal{A}(V^m_{t,i}) \simeq M_1(\mathbb{C}) \oplus M_1(\mathbb{C})$. By introducing appropriate commutation and anticommutation relations between the unitary selfadjoint generators of the "one-point" observable algebras (which relations respect microcausality) one can generate the net of local algebras. Since there is an increasing sequence of double cones covering $\mathcal{M}^2$ such that the corresponding local algebras are isomorphic to full matrix algebras $M_{2^n}(\mathbb{C})$, the quasilocal observable algebra $\mathcal{A}$ is a uniformly hyperfinite (UHF) C^*-algebra and consequently there exists a unique (non-degenerate) normalized trace $\mathrm{Tr}\colon \mathcal{A} \to \mathbb{C}$ on it.

Now, consider the double cones $V_A := V^m_{0,-1} \cup V^m_{\frac{1}{2},-\frac{1}{2}}$ and $V_B := V^m_{\frac{1}{2},\frac{1}{2}} \cup V^m_{0,1}$ and the "two-point" algebras $\mathcal{A}(V_A)$ and $\mathcal{A}(V_B)$ pertaining to them (see Fig. 2). It turns out that all the minimal projections in $A(\mathbf{a}) \in \mathcal{A}(V_A)$ and $B(\mathbf{b}) \in \mathcal{A}(V_B)$ can be parametrized by unit vectors $\mathbf{a}$ and $\mathbf{b}$, respectively in $\mathbb{R}^3$. Now, consider two

projections $A_m := A(\mathbf{a^m}); m = 1,2$ localized in V_A, and two other projections $B_n := B(\mathbf{b^n}); n = 1,2$ localized in the spacelike separated double cone V_B.

Let the state of the system be the singlet state ϕ^s defined in an appropriate way (by a density operator composed of specific combinations of generators taken from various "one-point" algebras). It turns out that in state ϕ^s the correlation between A_m and B_n will the one familiar from the EPR situation:

$$corr(A_m, B_n) := \phi^s(A_m B_n) - \phi^s(A_m)\,\phi^s(B_n) = -\frac{1}{4}\,\langle \mathbf{a^m}, \mathbf{b^n} \rangle \tag{13}$$

where $\langle\ ,\ \rangle$ is the scalar product in $\mathbb{R}^3$. In other words A_m and B_n will correlate whenever $\mathbf{a^m}$ and $\mathbf{b^n}$ are not orthogonal. To violate the Bell inequality (11) set $\mathbf{a^m}$ and $\mathbf{b^n}$ as follows:

$$\mathbf{a^1} = (0,1,0) \tag{14}$$

$$\mathbf{a^2} = (1,0,0) \tag{15}$$

$$\mathbf{b^1} = \frac{1}{\sqrt{2}}(1,1,0) \tag{16}$$

$$\mathbf{b^2} = \frac{1}{\sqrt{2}}(-1,1,0) \tag{17}$$

With this setting (11) will be violated at the lower bound since

$$\phi^s\big(A_1B_1 + A_1B_2 + A_2B_1 - A_2B_2 - A_1 - B_1\big) =$$

$$-\frac{1}{2} - \frac{1}{4}\left(\langle \mathbf{a^1}, \mathbf{b^1} \rangle + \langle \mathbf{a^1}, \mathbf{b^2} \rangle + \langle \mathbf{a^2}, \mathbf{b^1} \rangle - \langle \mathbf{a^2}, \mathbf{b^2} \rangle\right) = -\frac{1+\sqrt{2}}{2} \tag{18}$$

Now, the question as to whether the four correlations $\{(A_m, B_n); m,n = 1,2\}$ violating the Bell inequality (11) have a joint common causal explanation was answered in Hofer-Szabó and Vecsernyés (2012b) by the following

Proposition 3. *Let* $A_m := A(\mathbf{a^m}) \in \mathcal{A}(V_A)$, $B_n := B(\mathbf{b^n}) \in \mathcal{A}(V_B); m,n = 1,2$ *be four projections parametrized by the unit vectors* via *(14)–(17) violating the Bell inequality in the sense of (18). Then there exist a noncommuting join common cause* $\{C, C^\perp\}$ *of the correlations* $\{(A_m, B_n); m,n = 1,2\}$ *localizable in the common past* $V_C := V_{0,-\frac{1}{2};0,\frac{1}{2}}$ *of* V_A *and* V_B *(see Fig. 3).*

Observe that C is localized in the *common past* of the four correlating events that is in the region which seems to be the "physically most intuitive" localization of the common cause.

Propositions 2 and 3 together show that the relation between the common causal explanation and the Bell inequality in the noncommutative case is different from that in the commutative case. In the latter case the satisfaction of the Bell inequality is a necessary condition for a set of correlations to have a joint common

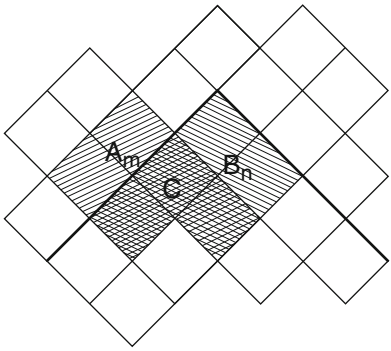

Fig. 3 Localization of a common cause for the correlations $\{(A_m, B_n)\}$

causal explanation. In the noncommutative case, however, the violation of the Bell inequality for a given set of correlations does *not* exclude the possibility of a joint common causal explanation for the set. And indeed, as Proposition 3 shows, one can find a common cause even for a set of correlations violating the Bell inequality. To sum it up, taking seriously the noncommutative character of AQFT where events are represented by not necessarily commuting projections, one can provide a common causal explanation in a much wider range than simply sticking to commutative common causes.

4 Conclusions

In the paper we were arguing that embracing noncommuting common causes in our explanatory framework is in line with the spirit of quantum theory and it gives us extra freedom in the search of common causes for correlations. Specifically, it helps to maintain the validity of Reichenbach's Common Causal Principle in the context of AQFT and it also helps to provide a local, *joint* common causal explanation for a set of correlations even if they violate the Bell inequalities.

Using noncommuting common causes naively to address the basic problems of the causal explanation in quantum theory in a formal way is no use whatsoever, if it is not underpinned by a viable ontology on which the causal theory can be based. This is a grandious research project. I conclude here simply by posing the central question of such a project:

Question. What ontology exactly is forced upon us by using noncommuting common causes in our causal explanation?

Acknowledgements This work has been supported by the Hungarian Scientific Research Fund OTKA K-100715.

References

Butterfield, J. 1989. A spacetime approach to the bell inequality. In *Philosophical consequences of quantum theory*, ed. J. Cushing and E. McMullin, 114–144. Notre Dame: University of Notre Dame Press.

Butterfield, J. 1995. Vacuum correlations and outcome independence in algebraic quantum field theory. In *Fundamental problems in quantum theory, annals of the New York Academy of Sciences, proceedings of a conference in honour of John Wheeler*, ed. D. Greenberger and A. Zeilinger, 768–785.

Butterfield, J. 2007. Stochastic Einstein locality revisited. *The British Journal for the Philosophy of Science* 58: 805–867.

Haag, R. 1992. *Local quantum physics*. Berlin: Springer.

Halvorson, H. 2007. Algebraic quantum field theory. In *Philosophy of physics*, vol. I, ed. J. Butterfield and J. Earman, 731–922. Amsterdam: Elsevier.

Halvorson, H., and R. Clifton. 2000. Generic bell correlation between arbitrary local algebras in quantum field theory. *Journal of Mathematical Physics* 41: 1711–1717.

Hofer-Szabó, G., and P. Vecsernyés. 2012a. Reichenbach's common cause principle in algebraic quantum field theory with locally finite degrees of freedom. *Foundations of Physics* 42: 241–255.

Hofer-Szabó, G., and P. Vecsernyés. 2012b. Noncommuting local common causes for correlations violating the Clauser–Horne inequality. *Journal of Mathematical Physics* 53: 12230.

Hofer-Szabó, G., and P. Vecsernyés. 2013a, submitted. Noncommutative common cause principles in algebraic quantum field theory. *Journal of Mathematical Physics* 54: 042301.

Hofer-Szabó, G., and P. Vecsernyés. 2013b, submitted. Bell inequality and common causal explanation in algebraic quantum field theory. *Studies in the History and Philosophy of Modern Physics* 44(4): 404–416.

Müller, V.F., and P. Vecsernyés. The phase structure of G-spin models, to be published.

Nill, F., and K. Szlachányi. 1997. Quantum chains of Hopf algebras with quantum double cosymmetry. *Communications in Mathematical Physics* 187: 159–200.

Rédei, M. 1997. Reichenbach's common cause principle and quantum field theory. *Foundations of Physics* 27: 1309–1321.

Rédei, M. 1998. *Quantum logic in algebraic approach*. Dordrecht: Kluwer.

Rédei, M., and J.S. Summers. 2002. Local primitive causality and the common cause principle in quantum field theory. *Foundations of Physics* 32: 335–355.

Rédei, M., and J.S. Summers. 2007. Remarks on causality in relativistic quantum field theory. *International Journal of Theoretical Physics* 46: 2053–2062.

Reichenbach, H. 1956. *The direction of time*. Los Angeles: University of California Press.

Summers, J.S. 1990. On the independence of local algebras in quantum field theory. *Reviews in Mathematical Physics* 2: 201–247.

Summers, J.S., and R. Werner. 1987a. Bell's inequalities and quantum field theory, I: General setting. *Journal of Mathematical Physics* 28: 2440–2447.

Summers, J.S., and R. Werner. 1987b. Bell's inequalities and quantum field theory, II: Bell's inequalities are maximally violated in the vacuum. *Journal of Mathematical Physics* 28: 2448–2456.

Summers, J.S., and R. Werner. 1988. Maximal violation of Bell's inequalities for algebras of observables in tangent spacetime regions. *Annales de l'Institut Henri Poincaré – Physique Théorique* 49: 215–243.

Szlachányi, K., and P. Vecsernyés. 1993. Quantum symmetry and braid group statistics in G-spin models. *Communications in Mathematical Physics* 156: 127–168.

Wald, R.M. 1992. Correlations beyond the horizon. *General Relativity and Gravitation* 24: 1111–1116.

Lost in Translation: A Comment on "Noncommutative Causality in Algebraic Quantum Field Theory"

Dustin Lazarovici

Let me guess. He pulled a lost in translation on you?

– from the movie *Ocean's Twelve* (2004)

What is the meaning of Bell's theorem? What are its implications for which it was dubbed, and rightfully so, "the most profound discovery of science" (Stapp 1975)? In brief, Bell's theorem tells us that certain statistical correlations between space-like separated events that are predicted by quantum mechanics and observed in experiment imply *that our world is non-local*. More precisely, it tells us that those correlations are *not locally explainable*, meaning that they cannot be accounted for by any local candidate theory since the frequencies predicted by a local account would have to satisfy a certain inequality – the Bell, respectively the CHSH inequality – that is empirically violated in the pertinent scenarios. Every candidate theory that correctly predicts the violation of the Bell-inequality must therefore describe some sort of direct influence between the correlating events, even if they are so far apart that they cannot be connected by a signal moving at maximum the speed of light. Hence, we say that the principle of *locality* or *local causality* is violated in nature.

The genius of Bell's argument lies in its simplicity and its generality. Bell's theorem is not about quantum mechanics, or quantum field theory, or *any* theory in particular, it is not confined to the "classical" domain or the quantum domain or a relativistic or non-relativistic domain, it is a *meta-theoretical* claim, excluding (almost) all possibilities of a local explanation for the statistical correlations

D. Lazarovici (✉)
Department of Mathematics, Ludwig-Maximilians-University of Munich,
Theresienstr. 39, 80333 Munich, Germany
e-mail: dustin.lazarovici@math.lmu.de

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_39,

observed in the so-called EPR-experiments.[1] Admittedly, a statement about nature can never reach the same degree of rigor as a theorem of pure mathematics for there is always an issue of connecting formal concepts to "real-world" concepts. Bell, however, was one of the clearest thinkers of the twentieth century and his analysis, unobscured by the misunderstandings of some of his later commentators, is absolutely precise and conclusive in this respect.[2] It is against this background that contributions to the subject have to be evaluated.

The paper *Noncommutative Causality in Algebraic Quantum Field Theory* by Gábor Hofer-Szabó, that I was gratefully given the opportunity to comment on, seems to be an offspring of a research project started about one and a half decades ago by Miklos Rédei (1997) and concerned with the question whether in Algebraic Quantum Field Theory[3] correlations between space-like separated events (in particular such violating the Bell-inequality) have local explanations in terms of "common causes". You see, what worries me about this research program is that it seems to suggest that the status of Bell's theorem is not yet clear, that the issue of non-locality is not yet settled, because somehow the technical details of Algebraic Quantum Field Theory could turn out to matter, and type III von-Neumann algebras could turn out to matter, and noncommutativity could turn out to matter. But that would be incorrect; none of this really matters.

Contrary to what I've just so emphatically stated, Mr. Hofer-Szabó makes quite an astonishing announcement. He claims that by committing ourselves to the framework of AQFT and by "embracing noncommuting common causes" we can achieve what Bell's theorem would seem to exclude, namely to provide a "local (joint common causal) explanation for a set of correlations violating the Bell inequalities". Although such a statement will certainly make a huge impression on people who believe that noncommutativity holds the one great mystery of quantum physics, we should pause for an instant to assess its plausibility.

---

[1] The only additional assumption is that the local candidate theories excluded by the theorem are supposed to be *non-conspiratorial* or *not superdeterministic*, meaning that certain parameter-choices involved in the experiments are assumed to be "free" or random and not predetermined in precisely such a way as to arrange for apparently non-local correlations. See Bell (1990) for details.

[2] A beautiful presentation of his analysis can be found in Bell (1981, 1990), the original version of the theorem is Bell (1964). For a more recent discussion, see Goldstein et al. (2011) or Maudlin (2011). The most common misunderstandings are addressed, for instance, in Norsen (2006) or Goldstein et al. (2011).

[3] Algebraic Quantum Field Theory is sometimes referred to as Local Quantum Field theory, but that is due to an unfortunate double-use of terminology. "Locality" in quantum field theory usually refers to the postulate of "microcausality" or "local commutativity" requiring that operators localized in space-like separated regions of space-time commute. This, however, is very different from the concept of Bell-locality as discussed above. In AQFT, local commutativity assures the impossibility of faster-than-light signaling, the theory nevertheless contains *non-local* correlations between space-like separated events due to the non-local nature of the quantum state (or most "states"), which is defined as a functional on the entire "net" of operator algebras all over space-time.

Physical events, or "causes" and "effects", whatever we might mean by that, are certainly not the kind of thing that can either commute or not commute. Operators, I grant, can commute or not commute, and so can perhaps elements of lattices with respect to certain set-theoretic operations. But Bell's theorem doesn't care about any of this. His argument is only concerned with the predictions that a candidate theory makes for the probabilities of certain events, not with the mathematical structure that it posits to make those predictions or represent those events. So how could it be possible to avoid the consequences of Bell's theorem by denying "commutativity", which hasn't been among its premises in the first place?

Let me try to explain what I think the result presented in the paper actually consists in and why it is missing the point as concerning the issue of local causality. Contrary to what is being suggested in the paper, the existence of a "commuting/noncommuting (weak/strong) (joint) common cause system" according to its definitions 2 and 3 is not sufficient for a local (common causal) explanation of correlations between space-like separated events. Such an explanation would be at least required to *reproduce* the statistical correlations that it was set out to explain. The kind of "explanation" that the author provides *doesn't do this*.

As his paper correctly states, the statistics for the events A_i, B_i are different whether the state is first projected on the possible "common causes" (since that's what happens when we compute $\phi \circ E_c$), or not. Most notably, the probabilities for the correlated events after the "occurrence" (more correctly: measurement) of "noncommuting common causes" (the right-hand-side of (1) below) satisfy the Bell inequality – in accordance with Bell's theorem – whereas the statistical correlations that the author *claims to explain* violate Bell's inequality. Note that, in the case where A_i, B_j don't commute with C_k, we will generally find that

$$\phi(AB) \neq \sum_k \frac{\phi(C_k A B C_k)}{\phi(C_k)} \phi(C_k). \tag{1}$$

This is, I assume, a familiar fact (if you find the notation confusing, write $\langle\psi|C_k A_i C_k|\psi\rangle$ for $\phi(C_k A_i C_k)$, and so on). Also, there is nothing deep, or mysterious, or metaphysically interesting about it, if only we appreciate the fact that the right-hand-side of (1) does not describe the same physical situation in which the system remains undisturbed in the common past of A and B, but that the projection on the common cause system (indeed one could think of a "measurement" of an observable C with spectral decomposition $\{C_k\}$) affects (decoheres) the quantum state in a way that can influence subsequent (measurement-)events. Here, it will simply "destroy" the EPR-correlations, so that violations of the Bell-inequality don't occur at all. In particular, the correlations described by the left-hand-side of (1) are *not explained* by the right-hand-side of (1), since the two probability distributions are different.

In Hofer-Szabó and Vecsernyés (2012), the author explicitly acknowledges this point, yet responds by saying that "the definition of the common cause does *not* contain the requirement (which our classically informed intuition would dictate) that

the conditional probabilities, when added up, should give back the unconditional probabilities [. . .]. Or, in other words, that the probability of the correlating events should be built up from a finer description of the situation provided by the common cause" (p. 20)

Although this doesn't strike me as a very strong argument, I think it's a good starting point to make a few remarks and highlight some of the disagreements between the author and myself.

1. As I see it, the problem here is not with probability theory ("that the conditional probabilities, when added up, should give back the unconditional probabilities"), but with the assumption that the "common cause system" provides a "finer description" of the *same* physical situation. The fact that $C + C^{\perp} = \mathbb{1}$ in terms of operators does not imply that it makes no difference whether any of the "events" occur, or none. There is a difference between a physical situation where a photon can either pass or not pass a polarization filter and a physical situation with no polarization filter at all.
2. Even if we accepted the premise of the answer, it wouldn't resolve the issue. The interesting question concerning correlations between space-like separated events is whether they can be explained by some sort of local "mechanism" (I'm using "mechanism" broadly here). Bell's theorem states this is impossible if the Bell inequality is violated. The fact that Hofer-Szabó presents us with a local mechanism that produces *different* correlations that do *not* violate Bell's inequality seems quite irrelevant in this context.
3. Despite the tone of the paper suggesting a certain naturalness or inevitability to the concepts it explores, we should keep in mind that it was the author himself who has chosen to *redefine* the "common cause principle" for the needs and purpose of AQFT (or rather has chosen to follow Rédei 1997, while admitting noncommuting operators). So when confronted with the objection that the concept he defined is unsubstantial because it lacks a certain crucial property, he cannot defend himself by pointing out that the concept lacks this property by definition. As far as I understand from this and former publications (e.g. Rédei 1997; Hofer-Szabó and Vecsernyés 2012) the reasoning behind their definition is that the common cause principle of Reichenbach is somehow "classical" and that there is a canonical way to "translate" or "generalize" it to AQFT. Leaving aside the question whether Reichenbach's common cause principle is at all a helpful concept in this context (since his discussion had a different focus), the statement that it is "classical" strikes me as rather confusing. Reichenbach's common cause principle is a philosophical concept, formulated in terms of what some people call "classical probability theory". But the word "classical" in "classical probability theory" shouldn't be confused with the same adjective in the term "Classical (i.e. Newtonian) Mechanics". It doesn't refer to a particular physical theory that can be empirically tested, but to a mathematical framework that expresses a certain *way of reasoning*. It is possible, of course, to take Reichenbach's definition and replace the probabilistic events, assumed

> to be modelled on a classical probability space, by projections in local algebras and the probability measure by a "state" that yields a value between 0 and 1 when evaluated on such projections; but that doesn't necessary mean that this procedure will yield a meaningful notion of "common causes" or "common cause explanations" in the context of *any* theory (regardless of the issue of commutativity). There are certainly people who believe that quantum theory is and has always been about replacing "classical probability spaces" by so-called "quantum probability spaces". However, I would think that for the sake of a meaningful philosophical discussion, we will have to do better. Anyway, I would insist that if we work under this general hypothesis, then the fact that the results we obtain are somehow counterintuitive or even logically inconsistent need not necessarily reflect some sort of quantum weirdness in nature; it may just as well reflect a lack of imagination or understanding on our side to appreciate that quantum physics is *not* always about putting little hats on capital A's and B's and C's to turn them into operators.[4]

In my opinion, much of the confusion in the paper stems from its commitment to a particular jargon that insists on using quite familiar terms ("events", "common causes", etc.) with a very non-standard meaning (usually referring to operators). Hence, I believe it is extremely illuminating to drop the jargon altogether and discuss the situation in good old-fashioned quantum mechanics to see what the result presented in the paper actually consists in.

Let's consider the usual spin-singlet state

$$\frac{1}{\sqrt{2}}\big(|\uparrow\rangle_1 \otimes |\downarrow\rangle_2 - |\downarrow\rangle_1 \otimes |\uparrow\rangle_2\big),$$

giving rise to EPR-correlations between the spin-orientations of two entangled electrons. These correlations can be observed between the results of spin-measurements performed simultaneously on both particles by Alice and Bob. Now, in the sense promoted in the paper, a "noncommuting joint common causal explanation" would, for instance, consist in the following: after the particles leave the EPR-source, we perform a z-spin measurement on particle 1, taking place in the common past of the measurements of Alice and Bob, who are free to choose between certain orientations other than the z-direction. This (chronologically) first measurement now provides a (non-trivial) "noncommuting joint common cause system" in the sense of Hofer-Szabó, namely

$$\big\{C = |\uparrow\rangle\langle\uparrow| \otimes \mathbb{1},\ C^{\perp} = |\downarrow\rangle\langle\downarrow| \otimes \mathbb{1}\big\}.$$

[4] About common controversies or misconceptions regarding the relevance of "quantum logic", "quantum probabilities" or "noncommutativity" for the issue of non-locality, see also Goldstein et al. (2011).

Obviously, the probabilities for the outcomes of the succeeding measurements will now split. Just as obviously, the first measurement will destroy (decohere) the singlet-state and the outcomes of the spin-measurements by Alice and Bob will not be correlated in a way that violates the Bell or CHSH inequality.

The result presented in the paper *Noncommutative Causality in Algebraic Quantum Field Theory*, although more general and technically more sophisticated, doesn't do anything more than this. I leave it to the reader to judge its explanatory value.

References

Bell, J.S. 1964. On the Einstein-Podolsky-Rosen paradox. Reprinted in Bell, J.S. 2004. *Speakable and unspeakable in quantum mechanics*, 14–21. Cambridge: Cambridge University Press.

Bell, J.S. 1981. Bertlmann's socks and the nature of reality. Reprinted in Bell, J.S. 2004. *Speakable and unspeakable in quantum mechanics*, 139–158. Cambridge: Cambridge University Press.

Bell, J.S. 1990. La Nouvelle cuisine. Reprinted in Bell, J.S. 2004. *Speakable and unspeakable in quantum mechanics*, 233–248. Cambridge: Cambridge University Press.

Goldstein, S., et al. 2011. Bell's theorem. *Scholarpedia* 6(10): 8378. http://www.scholarpedia.org/article/Bell's_theorem

Hofer-Szabó, G., and P. Vecsernyés. 2012, preprint. Bell inequality and common causal explanation in algebraic quantum field theory. http://philsci-archive.pitt.edu/9101

Maudlin, T. 2011. *Quantum non-locality and relativity*, 3rd ed. Malden/Oxford: Wiley-Blackwell.

Norsen, T. 2006. Bell locality and the nonlocal character of nature. *Foundations of Physics Letters* 19(7): 633–655.

Rédei, M. 1997. Reichenbach's common cause principle and quantum field theory. *Foundations of Physics* 27: 1309–1321.

Stapp, H.P. 1975. Bell's theorem and world process. *Il Nuovo Cimento* 40B(29): 270–276.

Causal Probabilities in GRW Quantum Mechanics

Tomasz Placek

1 Introduction

A large part of the debate over the interpretations of probabilities has been fueled by the Humean question as to whether there are irreducible modal factors (causal powers, dispositions, or propensities). A recent tendency is to investigate how competing probability interpretations handle probability ascriptions occurring in particular sciences. A case in point is the controversy concerning probabilities in the GRW version of quantum mechanics[1] as developed by Ghirardi, Rimini and Weber.[2] Thus, Frigg and Hoefer (2007) provided a (Lewisian) Humean Best System (HBS) analysis of GRW probabilities, in which modalities are reduced to a balance between some theoretical virtues; these authors claim HBS analysis is superior to a modally irreducible single-case approach to GRW theory. In sharp contrast to this approach, Dorato and Esfeld (2010) advocate the notion that GRW probabilities are best understood as irreducibly modal single-case propensities that measure the objects' power to localize. This stance is a part of these authors' larger project of conceiving GRW theory as a fundamental ontology of powers.

The present paper is intended as offering support for Dorato and Esfeld's project by addressing a certain formal problem. Our perception is that an intuition, based perhaps on scientists' announcements, that a theory has a modal character, falls short of being evidence for the theory having a modal character. Similarly, saying

[1] Another illustration of this tendency is the debate over probabilities in the consistent histories quantum mechanics; for an argument that these probabilities can be read as single case causal probabilities – see Müller (2007).

[2] See for instance Ghirardi et al. (1986).

T. Placek (✉)
Institute of Philosophy, Jagiellonian University, Grodzka 52, 31-044 Krakow, Poland
e-mail: Tomasz.Placek@uj.edu.pl

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_40,

that a theory's probabilities can be read as single-case propensities is not enough. The present paper thus aims to first investigate the modal and spatiotemporal underpinnings of GRW theory (read as a theory of powers) in a way that allows for the identification of propensities with weighted well-localized modalities. Second, it addresses a popular formal objection to propensities by constructing, in the spirit of Müller's (2005) causal probability spaces, the required Boolean algebras on which the propensities are defined.

We choose the flash ontology, as proposed by Bell (1987), with objects identified with galaxies of such flashes, which seems to be particularly handy in relativistic versions of GRW theory. The possible evolutions of the configurations of flashes are represented in a branching-time framework (for non-relativistic GRW) or in branching space-times (for relativistic GRW, Tumulka's (2006) version). Both frameworks yield a natural, though minimal, concept of transitions as pairs of events (in particular, flashes), such that the first is causally before the second. Then, following Müller's (2005) construction, a Kolmogorovian probability space is associated with each transition. As a result of modal and spatiotemporal constraints, the probability of a transition of arbitrary length is a function of probabilities of the basic transitions it involves. Importantly, the function allows for a failure of factorizability, leaving room for quantum non-locality.

The conclusion is that all three ingredients of GRW theory, understood as an ontology of powers, can be rigorously represented: irreducible modality, spatiotemporal aspects, and propensities as weighted modalities. Yet there remains a worry about bearers of powers (or of dispositions). Typically this role is played by enduring objects, which are absent from the flash ontology. Could then powers be powers of flashes?

The paper is organized as follows. The next Sect. 2 offers an outline of GRW theory. Its Sect. 2.1 sketches Tumulka's (2006) relativistic model of the GRW theory. Then Sect. 3 describes Dorato and Esfeld's case for conceiving GRW theory as an ontology of powers (or of depositions). Section 4 constructs a modal, branching-style structure, to account for irreducible modalities claimed by the above authors. The next Sect. 5 constructs Müller-style causal probability spaces as a basis for single-case propensities, advocated by the same authors. The construction is presented for the non-relativistic case and sketched for a relativistic case.

2 An Outline of GRW Quantum Mechanics

The aim of GRW theory is to resolve the measurement problem of quantum mechanics. In contrast to the standard picture of quantum mechanics, chancy occurrences are not restricted to the interactions between measuring apparatuses and measured objects. The chancy occurrences (called jumps, hits, or flashes) are a part and parcel of the evolution of quantum systems; to accommodate them, the Schrödinger equation is modified appropriately. Since we are mostly interested in the modal aspects of GRW theory, we here focus on how chancy occurrences are

thought to occur. We may think of the evolution of a system as involving a sequence of stages in which Nature selects in a chancy way a value of a parameter from a range of possible values. The presentation in this subsection closely follows Allori et al. (2008).

For an N-particle system[3] we define a collapse operator $\Lambda_i(x)$ on the Hilbert space for this system, for any point $x \in \mathbb{R}^3$ and for any particle label $i = 1, 2, \ldots, N$:

$$\Lambda_i(x) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{(\hat{Q}_i - x)^2}{2\sigma^2}}, \tag{1}$$

where $\hat{Q}_i$ is the position operator of particle i and σ is a (new) constant of nature. The system's evolution can be then described as proceeding by the following stages:

1. Let us suppose that the initial state of this system at time t_o is ψ_{t_0}.
2. Then Nature randomly selects time interval ΔT_1, the distribution of which is exponential, with rate $N\lambda$, where λ is another (new) constant of nature. Until time $T_1 = t_0 + \Delta T_1$ the system evolves unitarily, $\psi_0 \rightarrow \psi_{T_1} = U_{\Delta T_1}\psi_{t_0}$, where $U_{\Delta t_1}$ is the unitary operator corresponding the system's Hamiltonian.
3. At time T_1 the system undergoes a spontaneous hit, affected by two "chancy" choices on the part of Nature. First Nature selects a particle to be hit, i.e., it chooses a label I_1 out of N labels $1, 2, \ldots, N$, where a distribution is assumed to be uniform. Second, it selects a center of collapse, i.e., a location X_1, with the probability distribution

$$P(X_1 \in dx_1 \mid \psi_{T_1}, I_1 = i_1) = || \Lambda_{i_1}^{1/2}\psi_{T_1} ||^2 \, dx_1. \tag{2}$$

   As a result of the hit, the wave function transforms as follows:

$$\psi_{T_1} \rightarrow \psi_{T_1+} = \frac{\Lambda_{I_1}(X_1)^{1/2}\psi_{T_1}}{|| \Lambda_{I_1}(X_1)^{1/2}\psi_{T_1} ||} \tag{3}$$

4. The algorithm is iterated: Nature randomly selects time interval Δ_{T_2} (with exponential distribution). From T_1 to $T_2 = T_1 + \Delta_{T_2}$ the state ψ_{T_1+} evolves unitarily into ψ_{T_2}. Then Nature selects particle label I_2 and a center of the collapse X_2 (each with an appropriate distribution); at T_2, as a result of a hit, the state transforms as $\psi_{T_2} \rightarrow \psi_{T_2+}$, in accord with the analogue of Eq. 3.

Importantly, GRW theory yields the statistics of possible evolutions, given the initial state of the system. The probability distribution that after time t_0 n hits occur

[3] As GRW theory does not posit particles, "N-particles system" is a misnomer. If the theory is supplemented with a flash ontology, the phrase means that flashes come in kinds, and the system in question involves exactly N kinds of flashes.

at specified times $T_1, \ldots, T_n$ and locations $X_1, \ldots, X_n$, respectively, conditional on the system being in the initial state ψ_0, is the following:

$$P(X_1 \in dx_1, T_1 \in dt_1, I_1 = i_1, \ldots, X_n \in dx_n, T_n \in dt_n, I_n = i_n \mid \psi_{t_0}) = \lambda^n e^{-N\lambda(t_n - t_0)} \mid\mid L^{f_n}_{t_n,t_0} \psi_{t_o} \mid\mid dx_1 dt_1, \ldots, dx_n dt_n, \tag{4}$$

where

$$f_n = \langle\langle x_1, t_1, i_1\rangle, \ldots \langle x_n, t_n, i_n\rangle\rangle \tag{5}$$

and

$$L^{f_n}_{t_n,t_0} = U_{t_n - T_N} \Lambda_{I_n}(X_n)^{1/2} U_{T_n - T_{n-1}} \Lambda_{I_{n-1}}(X_{n-1})^{1/2} U_{T_{n-1} - T_{n-2}} \cdots \ldots \Lambda_{I_1}(X_1)^{1/2} U_{T_1 - t_0} \tag{6}$$

Equation 4 easily yields a formula for calculating a joint probability distribution for any later sequence of n "positions", where the positions are the system's states defined in terms of number of flashes together with their spatiotemporal locations. Let the initial position be the state immediately after m flashes, located at $X_1 T_1, \ldots, X_m, T_m$, whereas the final position is the state immediately after $m + n$ flashes, with the last n flashes located at $X_{m+1} T_{m+1}, \ldots, X_{m+n}, T_{m+n}$.

$$P(X_{m+1} \in dx_{m+1}, T_{m+1} \in dt_{m+1}, I_{m+1} = i_{m+1}, \ldots \ldots, X_{m+n} \in dx_{m+n}, T_{m+n} \in dt_{m+n}, I_{m+n} = i_{m+n} \mid \psi_{T_m +}) = = \lambda^n e^{-N\lambda(t_{m+n} - T_m)} \mid\mid L^{f_{m+n}}_{t_{m+n},t_m} \psi_{T_m +} \mid\mid dx_{m+1} dt_{m+1}, \ldots, dx_{m+n} dt_{m+n}, \tag{7}$$

with f_{m+n} and $L^{f_{m+n}}_{t_{m+n},t_m}$ defined by analogues of Eqs. 5 and 6, respectively, and $\psi_{T_m +}$ given by

$$\psi_{T_m +} = \frac{L^{f_m}_{t_m,t_0} \psi_{t_0}}{\mid\mid L^{f_m}_{t_m,t_0} \psi_{t_0} \mid\mid} \tag{8}$$

By summing over kinds of intermediate flashes, and integrating over positions and times of intermediate flashes, one gets a probability distribution of a transition from one flash to a later flash, given that the system's evolution started with the wave function ψ_{t_0}:

$$P(X_{m+n}, T_{m+n}, I_{m+n} \mid X_{m+1}, T_{m+1}, I_{m+1} \psi_{t_0}), \tag{9}$$

where X_{m+1}, T_{m+1}, and I_{m+1} abbreviate, respectively, $X_{m+1} \in dx_{m+1}, T_{m+1} \in dt_{m+1}$, and $I_{m+1} = i_{m+1}$.

Since between the flashes the system evolves unitarily, Eq. 7 yields a probability distribution for any transition, that is, not restricted to transitions between positions occurring *immediately* after flashes. We may as well consider transitions between positions like "the state Δ units of time after the last flash of a series of m flashes occurring at such-and-such spatiotemporal locations".

2.1 Tumulka's Relativistic Version of GRW

GRW theory can be supplemented by two ontologies: the flash ontology and the mass density ontology. The latter is in conflict with the spirit of relativity as a hit at one location involves an instantaneous change in mass density in remote regions.[4] Tumulka's (2006) relativistic model of GRW theory assumes the flash ontology; on this ontology, put forward by Bell (1987), spatiotemporal objects are identified with clouds of point-like flashes. This is a model for N non-interacting non-massless "particles", where the mention of particles should be understood as above, i.e., as the flashes coming in N varieties.

The system's wave function is defined on the product of N manifolds, $\prod_i^N \mathcal{M}_i$. Its evolution is governed by the Dirac equation plus the law of flashes. The law deriving the probability distribution of "new" flashes from the location of "old" flashes is constructed as follows. First, we introduce hyperboloids $\mathcal{H}_r(x) = \{y \in \mathbb{R}^4 \mid \text{t-dist}(x, y) = r\}$, where t-dist$(x, y)$ is the supremum of lengths of all curves connecting x and y. Let $X_1, X_2, \ldots, X_N$ be the "old" flashes of the kinds $1, 2, \ldots, N$, respectively, and ψ be the wave function of the system after these flashes.

We suppose that Nature selects N time intervals $\Delta T_1, \ldots, \Delta T_N$, which are mutually independent and each has exponential distribution with expectation τ. Together with "old" collapse centers $X_1, X_2, \ldots, X_N$, these intervals define N hyperboloids. We next assume that Nature selects N "new" collapse centers $Y_1, \ldots, Y_N$, each belonging to the appropriate hyperboloid, i.e., $Y_i \in \mathcal{H}_{c\Delta T_i}(X_i)$. The probability distribution of the transition from the "old" configuration of flashes to the "new" configuration of flashes is then calculated as

$$P(Y_1 \in d^3y_1, \ldots Y_N \in d^3y_N \mid X_1, \ldots, X_N, \psi) = \rho(y_1, \ldots, y_N) d^3y_1, \ldots, d^3y_n, \tag{10}$$

where d^3y_i is calculated by using Riemann metric on $\Sigma_i := \mathcal{H}_{c\Delta T_i}(X_i)$,

$$\rho(y_1, \ldots, y_N) = \int_{\prod_i \Sigma_i} d^3z_1 \ldots d^3z_N \mid j_{\Sigma_1}(y_1, z_1) \ldots j_{\Sigma_N}(y_N, z_N) \psi(z_1, \ldots, z_N)|^2, \tag{11}$$

[4] For more on the conceptual issues arising from combining GRW theory and relativity, see Maudlin (2008).

j_{Σ_i} are the jump factors defined by

$$j_\Sigma(y,z) = K_\Sigma(z) \exp \frac{\text{s-dist}^2_\Sigma(y,z)}{2a^2}. \tag{12}$$

s-dist(y,z) is infimum of lengths of curves joining y and z and $K_\Sigma(z)$ is chosen so that

$$\int_\Sigma d^3y \mid j_\Sigma(y,z) \mid^2 = 1. \tag{13}$$

The "new" wave function on the product $\prod \Sigma_i$ of hyperboloids is defined by

$$\varphi(z_1,\ldots,z_n) = \frac{j_{\Sigma_1}(Y_1,z_1)\ldots j_{\Sigma_N}(Y_N,z_N)\psi(z_1,\ldots,z_N)}{\rho^{1/2}(Y_1,\ldots,Y_N)} \tag{14}$$

and extended for the whole manifold product $\prod_i \mathcal{M}_i$ by solving the Dirac equation.

3 GRW as Ontology of Causal Powers

Dorato and Esfeld (2010) advocate conceiving of GRW theory as an ontology of causal powers, or of dispositions. Before we turn to the commitments this advocacy involves, we need to reflect on terminology. The above authors see a modal aspect of GRW theory in "dispositions for *spontaneous* localization". This phrase sounds, however, as if it were pushing in opposite directions: the word "disposition" has deterministic underpinning, whereas "spontaneous" suggests chanciness, randomness, or indeterminism. After all, a typical analysis of dispositions begins with this condition-manifestation schema:

> x has a disposition for F means that if x were subjected to condition C, a manifestation M would occur.

It is debatable, however, whether the schema handles indeterministic happenings. To illustrate, an electron's tendency to be deflected in a particular direction in the Stern-Gerlach experiment is not based on there being some condition C that ensures the deflection in this direction. The condition-manifestation condition pushes one to produce a deterministic hidden-variable model of quantum phenomena. That kind of observation has led some researchers, in the context of Bell's theorem, to draw a distinction between deterministic dispositions and indeterministic dispositions. For an analysis of the latter, the schema above should be amended – a typical suggestion is to add probabilities.

Instead of monstrous "indeterministic dispositions", I would use "powers". Powers are anchored in concrete objects (particles, fields). Powers are modal, and require the so-called real possibilities (aka historical possibilities). An object's

power rules out what is possible, and what is not for this object. Finally, an inherent part of powers is the gradation of possibilities: in our case, a power anchored in a configuration of flashes dictates the degrees of possibility of later configurations of flashes.

In the rest of this section we list the commitments inherent to the project of conceiving GRW theory as an ontology of causal powers, by quoting relevant passages from the Dorato and Esfeld's paper.

1. Irreducibility of causal powers (dispositions):

> [...] GRW admits events of spontaneous localization [...]. [N]on-massless microsystems possess a disposition for spontaneous localization. [T]his disposition is irreducible: it is not grounded on non-dispositional, categorical properties. It belongs to the ontological ground floor, so to speak, and it is a real and actual property, not a purely possible property. It is therefore appropriate to talk in terms of a power for spontaneous localization [...]

2. GRW probabilities are propensities:

> [...] [A]n important advantage of the propensity interpretation of probabilities is the possibility to attribute single-case probabilities. In order to talk about an objective, mind-independent probability, as GRW requires, a propensity theorist has the advantage of not having to refer to ensembles of particles, or to actual or idealized frequencies. Frequencies are simply supervenient on, and a manifestation of, those propensities to localize that are the essence of spatially superposed quantum particles.

3. Primacy of capacities over laws; and
4. modal (rather than Humean Best System) analysis of probabilities:

> [G]iven our commitment to the claim [...] that law statements are made true by dispositions or causal powers possessed by physical systems, we rely on the view that it is capacities rather than laws that are basic [...]. Consequently, it seems to us that an objectivist view of the GRW probabilities can be more naturally defended by committing oneself to mind-independent properties or relations that microsystems have, and by conceiving these properties or relations in a modal way, that is, as dispositions or causal powers. The HBS position, on the contrary, seems to oscillate between frequentism, with all its known problems, and epistemic views of probability, introduced by the criteria of simplicity and strength of laws, pushing one's position toward subjectivism or Bayesianism.

3.1 Modal Analysis of Probabilities

The project of reading GRW theory as an ontology of causal powers calls for a development of two themes. First, since modality is assumed to be irreducible, we need to supplement the theory with some modal structure. For five decades, the standard tool for analyzing modalities are possible worlds theories. Since the GRW scenarios are intuitively seen as growing out of a single trunk (specified by an initial wave function) we will use a branching theory, in the spirit of Prior-Thomason branching time, or Belnap's branching space-times. Second, since modal

powers need to be anchored in objects, one needs to develop a modally-thick notion of objects, the notion that conceives of objects as having alternative future developments, some of which become actualized in due course, and some of which do not. I will not be concerned here with this admittedly harder task. Finally, having a branching framework in place, it needs to be seen that the GRW probabilities are indeed weighted possibilities. This last task involves addressing a long-standing objection to propensities which says that although proponents of propensities identify probabilities and weighted possibilities, there are no algebraico-modal structures that justify this move (recall that classical probability is a measure on a Boolean algebra).

4 Modality First

In describing an algorithm for the generation of subsequent flashes as well as the changes in the wave function I used the anthropomorphism "Nature selects" to indicate how a system's alternative scenarios should be thought of. In the standard, i.e., non-relativistic case, a scenario can be represented as a linear dense continuous ordering, with its elements identified with simultaneity slices, together with the information as to when, where, and what kind of flashes occur. In what follows we will arrange the plethora of the scenarios in a tree with a single trunk, arriving at a branching-time (BT) model.

Before we proceed, we need some rudimentary information on BT models.

Definition 1. A BT model is a non-empty partial order $\mathcal{W} = \langle W, \leqslant \rangle$, with no backward branching, i.e.,

$$\forall e, e', e'' \in W \ (e' \leqslant e \wedge e'' \leqslant e \rightarrow (e' \leqslant e'' \vee e'' \leqslant e')),$$

and connections in the past, i.e.,

$$\forall e, e' \in W \ \exists e^* \in W \ (e^* \leqslant e \wedge e^* \leqslant e').$$

Elements of W, called "events", can be thought of as simultaneity slices in a Newtonian universe, and $x \leqslant y$ means that y is identical to x or belongs to a possible future of x. A history is defined as a maximal chain in $\mathcal{W}$. (Histories, as defined above, exist by the Zorn-Kuratowski lemma). A notion of alternative possibilities open at a given event can be constructed by first defining, for any $e \in W$, the relation $\equiv_e$ of undividedness on the set $H_{(e)}$ of histories containing e, that is,

$$\text{For all } h, h' \in H_{(e)} \ h \equiv_e h' \text{ iff } \exists e'(e' \in h \cap h' \wedge e < e') \tag{15}$$

It can be proved that $\equiv_e$ is an equivalence relation on $H_{(e)}$. It thus yields partition Π_e of $H_{(e)}$, which we call "the set of possibilities open at e".

To link non-relativistic GRW models to BT models we proceed as follows. We identify each scenario of a GRW system with the real line together with a code – (possibly infinite) sequence $c = \langle \psi_{t_0}, t_0, T_1, I_1, X_1, \ldots, T_n, I_n, X_n, \ldots \rangle$ such that ψ_{t_0} is a initial wave function at time $t_0, T_i \in \mathbb{R}$, $I_i = 1, \ldots, N$, $X_i \in \mathbb{R}^3$, and $t_0 < T_1 < T_2 < \ldots$. (The meaning of these symbols is explained in the algorithm described in Sect. 2.) It is assumed that all considered scenarios share the same t_0 and the same ψ_{t_0}.

Mathematically speaking, a scenario is thus a Cartesian product $\mathbb{R} \times \{c\}$, where c is a code. To obtain a branching structure of possible histories from scenarios, we need to paste scenarios in an appropriate way. Consider thus the following binary relation on the set $\mathcal{S}$ of all elements of all scenarios:

Definition 2. For $\langle x, c \rangle \in \mathbb{R} \times \{c\}$ and $\langle x', c' \rangle \in \mathbb{R} \times \{c'\}$, $\langle x, c \rangle \equiv \langle x', c' \rangle$ iff $x = x'$ and $c_{\leqslant x} = c'_{\leqslant x}$, where $c_{\leqslant x}$ is the initial segment of code c that does not contain any triple $\langle T_n, I_n, X_n \rangle$ with $x < T_n$.

It is easy to see that $\equiv$ is a reflexive, symmetric, and transitive relation on $\mathcal{S}$. Accordingly, the partition $W := \mathcal{S}/\equiv$ of $\mathcal{S}$ by $\equiv$ is well-defined. We will write elements of this structure as $[x, c] = \{\langle x', c' \rangle \mid \langle x, c \rangle \equiv \langle x', c' \rangle\}$. We define next a partial ordering on W by putting $[x, c] \leqslant [x', c']$ iff $x \leqslant x'$ and $[x, c] = [x, c']$. Accordingly, there are in $\langle W, \leqslant \rangle$ maximal chains, which we now call "possible histories". Furthermore, it is straightforward to convince ourselves that the model satisfies no backward branching and that any two elements of $\langle W, \leqslant \rangle$ have a lower bound in W, that is, $\langle W, \leqslant \rangle$ satisfies the requirement of connections in the past. We thus have a verdict that $\langle W, \leqslant \rangle$, a modal structure derived from the (non-relativistic) GRW theory, is a BT model. Moreover, in this model there is a maximal element in the intersection of every two possible histories. We note, however, that, since Nature can select a flash to occur at any location, the resulting BT model is large, comprising undenumerably many histories.

A similar construction can be carried out for the special-relativistic version of Tumulka's GRW model, the result being a branching space-times (BST) model. In this construction a scenario should be understood as a pair consisting of $\mathbb{R}^4$ together with a (possibly infinite) sequence coding the spatiotemporal locations and kinds of subsequent groups of flashes. Recall that in Tumulka's algorithm, flashes come in groups of N flashes, each flash belonging to a different kind. To achieve pasting of scenarios into BST possible histories, one needs to consider a relation of similarity between the points of any two scenarios. The idea is that two points, each belonging to a different scenario, are similar iff they have the same spatiotemporal coordinates and nothing in the past differentiates one from the other. The second conjunct naturally translates into a requirement on the codes of the two scenarios: the segment of the code representing flashes in the past of the first point should be identical to the segment of the code representing flashes in the past of the second point. The construction is nevertheless not straightforward, since to ensure the satisfaction of BST axioms, one needs to assume some additional topological postulates – this is needed for cases with infinitely many splitting points. Since Nature can select any location for flashes (which should be viewed as splitting points, given their modal

character), it is exactly the infinitary case that is needed here. We do not present it here, however, as similar constructions are carried out in Wroński and Placek (2009) and Placek and Belnap (2012).

5 Probability Next

5.1 Probabilistic Structures for the Non-relativistic GRW

Let us reflect upon where we are. On the one hand we have GRW formulas (Eqs. 9 and 10) that give (or can be used to calculate) a numerical value for the probability that a system undergoes a transition from one configuration of flashes to some other configuration of flashes. On the other hand, we have branching models (BT or BST) that provide modal underpinnings, in terms of sets of possible histories, to GRW theory. Emphatically, having a formula for probabilities does not suffice for taking these probabilities for graded modalities. And merely naming them "graded possibilities" will not do, either. We need to combine the mentioned formulas with a modal theory, here, of a branching variety.

In the non-relativistic case the task is to add probabilities to the tree-like picture sketched above. In this picture, alternative future possibilities available at some point are represented by branches of possible histories that extend upward from that point.

We thus have a natural candidate for a locus for probabilities in a BT model. Recall that a classical probability space is a triple $\langle A, F, \mu\rangle$, where A is a non-empty set (called master set), F is a σ-field on A, and μ is a normalized to unity measure on F. It is thus natural to associate with each $e \in W$ the following probability space: the master set A_e is identified with Π_e (the set of possibilities open at e), F_e is a σ-field on A, and μ_e is a normalized to unity measure on F_e.

Although this procedure will land us with a plethora of probability spaces $\langle \Pi_e.F_e, \mu_e\rangle$, each associated with some event e, and with many such probability spaces being trivial,[5] the picture still remains intuitive: each $e \in W$ has a set of alternative future possibilities and how strong the possibilities are is measured by μ_e. Despite its naturalness, as well as the ability to accommodate an indefinitely large set of open possibilities, the approach is too detailed to work in the case of GRW. Since the GRW algorithms assign probabilities to transitions from one configuration of flashes to another, to represent the GRW probabilities we need to combine measures from different probability spaces, say μ_e $\mu_{e'}$, $\mu_{e''}$. This is not an easy task, as it requires addressing questions of dependence (or independence) of probability measures. We thus favor a different approach, called causal probability

[5] By trivial case I mean here a case with Π_e having one element only.

spaces, that was developed in Müller (2005, 2011). We need to warn the reader, however, that this approach is not immediately applicable to the GRW probabilities, because of its twofold finitistic assumptions: it assumes that the number of choice points is finite and that each choice point has finitely many open future possibilities. We nevertheless present it as an adequate analysis of the GRW probabilities, modally understood, since the approach seems to be capable of being generalized to infinitary cases,[6] and we are here interested in conceptual analysis rather than mathematical details.

We will need the following notions (each notion can be used both in the branching time framework as well as the branching space-times frameworks; they were first introduced by Belnap (2003)):

1. *Initial event* I is a subset of a history;
2. *Outcome chain* O is a non-empty lower bounded chain of point events;
3. *Scattered outcome* $\mathbf{O}$ is a set of outcome chains all of which overlap some one history;
4. *Transition from initial event* I *to outcome chain* O, $I \rightarrowtail O$, is a pair: initial event I and outcome chain O such that $\forall e \in I \; e < O$, where $e < O$ means that $\forall x \in O \; e < x$;
5. *Transition from initial event* I *to scattered outcome* $\mathbf{O}$, $I \rightarrowtail \mathbf{O}$, is a pair: initial event I and scattered outcome event $\mathbf{O}$ such that

$$\forall e \; (e \in I \rightarrowtail \exists O \; O \in \mathbf{O} \wedge e < O);$$

6. Basic transition, written as $e \rightarrowtail H_e$, is a pair: a point event e and an elementary possibility H_e open at e, i.e., $H_e \in \Pi_e$;
7. A set $\{e_\alpha \rightarrowtail H_\alpha\}_{\alpha \in I}$ of basic transitions is consistent iff $\bigcap_{\alpha \in I} H_\alpha \neq \emptyset$;
8. Basic transitions can be ordered as follows: for basic transitions $t_1 = e_1 \rightarrowtail H_1$ and $t_2 = e_2 \rightarrowtail H_2$, $t_1 \preceq t_2$ iff $e_1 \leqslant e_2$ and $e_2 \in \bigcup H_1$. Provably $\preceq$ is a partial ordering on the set of basic transitions.

Given this background, Müller's (2005) construction proceeds as follows:

1. Begin with a set T of basic transitions;
2. Produce T^* by adding to T every basic transition that is alternative to some element of T;
3. Consider maximal consistent subsets of T^*; the set of maximal consistent subsets of T^* is the master set of a sought-for probability space.

Let us illustrate how one applies this instruction. Suppose we want to assign probability, $pr(I \rightarrowtail \mathbf{O})$, to transition $I \rightarrowtail \mathbf{O}$, so that we need to construct a probability space in which this probability could be properly represented. Observe first that in order to pass from initial I to outcome event $\mathbf{O}$, choice points for

[6] See Sylvia Wenmackers's work, http://www.sylviawenmackers.be

appropriate histories need to cooperate: at each relevant choice point a possibility should be selected that does not prohibit the occurrence of **O**. Consider thus choice points between, on the one hand, histories containing I and passing through **O**, and, on the other hand, histories containing I but altogether avoiding **O**. These choice are relevant for the transition from I to **O**. One requires further that they lie in the past of outcome **O**, we arrive at the following notion of *past causal loci* for transitions $I \rightarrowtail \mathbf{O}$, $pcl(I \rightarrowtail \mathbf{O})$, due to Belnap (2005):

$$pcl(I \rightarrowtail \mathbf{O}) \stackrel{df}{=} \{e \mid \exists O(e < O \wedge O \in \mathbf{O}) \wedge \exists h(h \in H_{[I]} \wedge h \perp_e H_{\langle \mathbf{O} \rangle})\} \quad (16)$$

Belnap defines the originating causes, or causae causantes $cc(I \rightarrowtail \mathbf{O})$, of transition $I \rightarrowtail \mathbf{O}$ as:

$$cc(I \rightarrowtail \mathbf{O}) \stackrel{df}{=} \{e \rightarrowtail \Pi_e\langle \mathbf{O} \rangle : e \in pcl(I \rightarrowtail \mathbf{O})\}, \quad (17)$$

where $\Pi_e\langle \mathbf{O} \rangle$ is a (unique) element $H \in \Pi_e$ such that for some $h \in H$ and every $O \in \mathbf{O} : h \cap O \neq \emptyset$.[7] A significant fact, proved in the same paper, is that, given the form of occurrence propositions for transitions, each causa causans of Eq. 17 is an INNS condition for $I \rightarrowtail \mathbf{O}$, i.e., an insufficient but non-redundant part of a necessary and sufficient condition for the occurrence of $I \rightarrowtail \mathbf{O}$.

Whatever is responsible for transition $I \rightarrowtail \mathbf{O}$ to come through, is located at $pcl(I \rightarrowtail \mathbf{O})$; moreover, the occurrence of all $cc(I \rightarrowtail \mathbf{O})$ guarantees $I \rightarrowtail \mathbf{O}$ to come through. All other events are inert in bringing about our transition because either they are not below outcome **O**, or are deterministic in the sense that have only one future possibility. It is thus natural to identify the probability of transition $I \rightarrow \mathbf{O}$ with the probability of the set $cc(I \rightarrow \mathbf{O})$ of causae causantes of $I \rightarrow \mathbf{O}$. Following the recipe above, we begin with the set $cc(I \rightarrow \mathbf{O})$; its elements are basic transitions. Then we complete it to $cc^*(I \rightarrow \mathbf{O})$ by adding alternatives to every element of it. Finally, we take as a master set the set $cc^{**}(I \rightarrow \mathbf{O})$ of all maximal chains in $cc^*(I \rightarrow \mathbf{O})$. We write $\mu_{cc(I \rightarrow \mathbf{O})}$ for a measure on the field of subsets of $cc^{**}(I \rightarrow \mathbf{O})$. (Note that the subscript refers here to the set of basic transitions we started with rather than to a resulting master set, or the field of subsets.) The idea that only causae causantes count in determining the probability of a transition receives now the following reading:

$$pr(I \rightarrow \mathbf{O}) = \mu_{cc(I \rightarrow \mathbf{O})}(cc(I \rightarrow \mathbf{O})) \quad (18)$$

Observe that indeed $cc(I \rightarrow \mathbf{O})$ is a proper argument for the measure function, as it is a maximal chain in $cc^*(I \rightarrow \mathbf{O})$, so it is an atom in the field on $cc^{**}(I \rightarrow \mathbf{O})$. Importantly, causal probability spaces provide a hospitable environment for

[7] For a proof that $\Pi_e\langle \mathbf{O} \rangle$ exists and is unique, see Belnap (2003).

correlations. In itself, the machinery does not enforce the probability of a transition is a product of probabilities of the component transitions. In the non-relativistic case, with the assumed representation of the world's temporal slices by points, correlations are hardly desirable, as they may only mean correlations between temporally ordered transitions. However, the capacity to represent correlation is advantageous in a relativistic context, as modeled by BST, where consistent transitions need not be comparable (by $\preceq$).

It remains to coin a link between the machinery of initial events, outcome events, and transitions, and the modal reading of the GRW theory, sketched in Sect. 4. First, we need to relate what we called "positions", i.e., triples T_n, X_n, I_n consisting of a time of a flash, its spatial location, and its kind, to elements of a branching (BT) tree. Note that the same position can occur in many branches, as the same flash may occur in different ways, each way having a different past. Since in a BT representation of the non-relativistic GRW, events are the equivalence classes $[t, c]$, where t is a time coordinate and c is a code, to determine an event one needs to consider a position, T_n, X_n, I_n together with an appropriate code, where "appropriate" means that the triple T_n, X_n, I_n is a segment of that code. We will thus denote an event obtained from a given position and an appropriate code as $[T_n, c_{T_n,X_n,I_n}]$, where c_{T_n,X_n,I_n} is a code c that contains, as a segment, the triple T_n, X_n, I_n. Clearly, each $[T_n, c_{T_n,X_n,I_n}]$ is an element of a possible history, and hence it may be thought of as an initial event as well as an outcome event. Finally, to characterize transitions, we need to reflect on the ordering relation. We have:

$[T_n, c_{T_n,X_n,I_n}] \leqslant [T_m, c_{T_m,Y_m,I_m}]$ iff $T_n \preceq T_m$ and there is a code c^* such that $[T_n, c_{T_n,X_n,I_n}] = [T_n, c^*_{T_n,X_n,I_n}]$ and $[T_m, c_{T_m,Y_m,I_m}] = [T_m, c^*_{T_m,Y_m,I_m}]$.

Thus, in accord with the definition of transitions, as given above, we write $[T_n, c_{T_n,X_n,I_n}] \rightarrowtail [T_m, c_{T_m,Y_m,I_m}]$ for a pair $\langle [T_n, c_{T_n,X_n,I_n}], [T_m, c_{T_m,Y_m,I_m}] \rangle$ such that $[T_n, c_{T_n,X_n,I_n}] \leqslant [T_m, c_{T_m,Y_m,I_m}]$. It remains to relate the "numerical" formula of Eq. 7 to the "conceptual" formula of Eq. 18:

$$\begin{aligned} pr([T_m, c_{T_m,X_m,I_m}] \rightarrowtail [T_{m+n}, c_{T_{m+n},Y_{m+n},I_{m+n}}]) = \\ P(X_{m+1} \in dx_{m+1}, T_{m+1} \in dt_{m+1}, I_{m+1} = i_{m+1}, \ldots \\ \ldots, X_{m+n} \in dx_{m+n}, T_{m+n} \in dt_{m+n}, I_{m+n} = i_{m+n} \mid \psi_{T_m+}), \end{aligned} \tag{19}$$

where P is the GRW probability of Eqs. 7 and 9, $X_{m+1}, T_{m+1}, I_{m+1}, \ldots X_{m+n}, T_{m+n}, I_{m+n}$ are such that every code c^* that witnesses that $[T_m, c_{T_m,X_m,I_m}]$ and $[T_{m+n}, c_{T_{m+n},Y_{m+n},I_{m+n}}]$ form a transition, has the same beginning segment, namely

$$\mathrm{Ini}_{m-1}, T_m, Y_m, I_m, T_{m+1}, X_{m+1}, I_{m+1}, \ldots T_{m+n}, X_{m+n}, I_{m+n},$$

where Ini_{m-1} is a fixed sequence of $m-1$ triples T, X, I.

5.2 *Probabilistic Structures for a Relativistic Version of GRW*

The branching space-time model is defined as a non-empty, partially ordered set $\mathcal{W} = \langle W, \leqslant \rangle$, which satisfies a few postulates.[8] The most important difference from branching time is the concept of history, since BST histories are defined as maximal upward directed subsets of a base set W rather than maximal chains in W. However, the definitions of events, chains, outcomes, and transitions stay in place, though the BST framework brings in some novel features of defined objects. Importantly, transitions occurring in a single history need not be linearly ordered, as the initials of two transitions may be space-like related. Also, a BST model can exhibit what Belnap (1992) calls "modal funny business"; it obtains if there are two basic transitions, $t_1 = e_1 \to H_1$ and $t_2 = e_2 \to H_2$ such that although the initials e_1 and e_2 belong to some one history, the outcomes do not overlap, i.e., $H_1 \cap H_2 = \emptyset$. Müller's (2005) construction of causal probability spaces in the framework of BST requires, apart from introducing the already mentioned finitistic assumptions, that a BST model does not contain modal funny business.

To each transition $I \rightarrowtail \mathbf{O}$ the construction associates a probability space produced out of the set of $cc(I \rightarrowtail \mathbf{O})$ of causae causantes for this transition; in this space probability of $I \rightarrowtail \mathbf{O}$ is representable as the probability of the set of all causae causantes for $I \rightarrowtail \mathbf{O}$. The theory has a built-in ability to represent non-local correlations, as the equality below need not hold,

$$\mu_{\{t_1,t_2\}}(\{t_1, t_2\}) = \mu_{\{t_1\}}(\{t_1\}) \times \mu_{\{t_2\}}(\{t_2\}).$$

where $t_1, t_2 \in cc(I \rightarrowtail \mathbf{O})$ for some transition $I \rightarrowtail \mathbf{O}$, and initials of t_1 and t_2 are space-like related.

To find a modal interpretation of probabilities delivered in Tumulka's non-relativistic version of GRW one needs to do two things: construct a BST model representing the possible developments of a system in Tumulka's model, and then link Tumulka's probabilities for the occurrence of a collection of configurations of flashes to BST probabilities, as defined on transitions between BST events. As we have already commented on the former task at the end of Sect. 4, we turn now to the latter.

Since a scenario in Tumulka's model can be viewed as a collection of sequences of nested hyperboloids, with each element of the collection associated with hits of a given *kind*, and each hyperboloid having a hit located on it, a scenario is specified by a code of the form

$$\begin{aligned} &T_1^1, X_1^1, I_1^1, T_2^1, X_2^1, I_2^1, \ldots, T_N^1, X_N^1, I_N^1, T_1^2, X_1^2, I_1^2, T_2^2, X_2^2, I_2^2, \ldots, T_N^2, X_N^2, I_N^2, \\ &\ldots, T_1^m, X_1^m, I_1^m, T_2^m, X_2^m, I_2^m, \ldots, T_N^m, X_N^m, I_N^m, \ldots, \end{aligned} \tag{20}$$

[8] For the postulates see, for instance Belnap (2003).

where superscripts refers to stages of configurations, whereas subscript differentiates between kinds of flashes. Then a BST event representing a configuration of flashes at a given stage is determined by giving a pair consisting of a sequence

$$T_1^m, X_1^m, I_1^m, T_2^m, X_2^m, I_2^m, \ldots, T_N^m, X_N^m, I_N^m$$

and a code, provided that the code contains the above sequence as its segment.

Observe finally that Eq. 10, by assigning probability to one configuration of flashes conditional on the last configuration of flashes, can be naturally read as the probability of transitions between subsequent (BST representations of) configuration of flashes. The equation can also be used to calculate the probability of transition between any two properly ordered BST events representing configurations of flashes.

6 Conclusions

We have provided a modally-temporal framework for the non-relativistic GRW theory and sketched a modally spatiotemporal framework for a special-relativistic version of this theory. The framework allows one to read indeterministic flashes in a modal and local way: at a given location (temporal or spatiotemporal) a flash may occur, but does not have to. Next, using this modal framework we provided a reading of GRW probabilities as graded possibilities. As far as providing formal foundations can support philosophical doctrines, this enterprise supports Dorato and Esfeld's position that GRW theory is naturally read as an ontology of causal powers. There is a caveat, however: typically, powers require bearers, they are powers *of* some objects to do something yet none of the candidates for a GRW ontology posits objects.

References

Allori, V., S. Goldstein, R. Tumulka, and N. Zanghì, 2008. On the common structure of Bohmian mechanics and the Ghirardi-Rimini-Weber theory. *The British Journal for the Philosophy of Science* 59(3): 353–389.

Bell, J. 1987. Are there quantum jumps? In *Schrödinger. Centenary of a polymath*. Cambridge: Cambridge University Press. Reprinted in Bell, J.S. 1987. *Speakable and unspeakable in quantum mechanics*. Cambridge: Cambridge University Press.

Belnap, N. 1992. Branching space-time. *Synthese* 92: 385–434.

Belnap, N. 2003. No-common-cause EPR-like funny business in branching space-times. *Philosophical Studies* 114: 199–221.

Belnap, N. 2005. A theory of causation: Causae causantes (originating causes) as inus conditions in branching space-times. *The British Journal for the Philosophy of Science* 56: 221–253.

Dorato, M., and M. Esfeld. 2010. GRW as an ontology of dispositions. *Studies in History and Philosophy of Modern Physics* 41(1): 41–49.

Frigg, R., and C. Hoefer. 2007. Probability in GRW theory. *Studies in History and Philosophy of Modern Physics* 38: 371–389.

Ghirardi, X., Y. Rimini, and Z. Weber. 1986. Unified dynamics for microscopic and macroscopic systems. *Physical Review D* 34: 470–491.

Maudlin, T. 2008. Non-local correlations in quantum theory: How the trick might be done. In *Einstein, relativity and absolute simultaneity*, ed. W.L. Craig and Q. Smith. London/New York: Routledge.

Müller, T. 2005. Probability theory and causation: A branching space-times analysis. *The British Journal for the Philosophy of Science* 56(3): 487–520.

Müller, T. 2007. Branch dependence in the 'consistent histories' approach to quantum mechanics. *Foundations of Physics* 37(2): 253–276.

Müller, T. 2011. Probabilities in branching structures. In *Explanation, prediction and confirmation*, ed. D. Dieks, W.J. Gonzalez, S. Hartmann, T. Uebel, and M. Weber, 109–121. Berlin: Springer.

Placek, T., and N. Belnap. 2012. Indeterminism is a modal notion: Branching spacetimes and Earman's pruning. *Synthese* 187(2): 441–469. doi:10.1007/s11229-010-9846-8.

Tumulka, R. 2006. A relativistic version of the Ghirardi-Rimini-Weber model. *Journal of Statistical Physics* 125(4): 825–844.

Wroński, L., and T. Placek. 2009. On Minkowskian branching structures. *Studies in History and Philosophy of Modern Physics* 40: 251–258.

Physics, Metaphysics and Mathematics

Dennis Dieks

1 Introduction

In some respects the history of philosophy of science is like a pendulum: whereas philosophers like Hume and Mach warned against mixing science and metaphysics, and in particular against the intrusion of causal concepts in physics, quite a number of modern philosophers of science argue that science and metaphysics are interdependent and that in fact science offers support for a causal metaphysics. Mach thought that the use of concepts like "force", when not explicitly made harmless by an analysis in empiricist terms, would introduce "animistic" connotations into physics and in this way veil the real scientific content – worse, such concepts could easily introduce pseudo-explanations that at bottom are incomprehensible or meaningless. This general line of reasoning was taken over by the logical positivists, and later by Humean philosophers like David Lewis (1986). But recent decades have seen a rebirth of traditional metaphysics in connection with science. An early example is the book "Causal Powers" by Harré and Madden (1975); more recently Bird's "Nature's Metaphysics" (Bird 2010) has appeared, in which a "dispositional essentialism" is defended according to which all fundamental properties have dispositional power-like essences. What these and other philosophers argue is that talk of causal powers, far from obscuring the real content of science, creates an understandable picture of what science is about. Good physical theories reveal that our world consists of causal powers, and conversely, amenability to a consistent causal interpretation can serve as a criterion for selecting plausible truth-candidates from a collection of rival theories.

D. Dieks (✉)
Institute for History and Foundations of Science, Utrecht University, P.O. Box 80.010, 3508 TA Utrecht, The Netherlands
e-mail: d.dieks@uu.nl

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_41,

Tomasz Placek's paper "Causal Probabilities in the GRW Quantum Mechanics" fits into this new metaphysically inspired tradition. The paper is part of a wider research program that seeks to detail and work out a specific conception of time and of modality. According to this conception, there is a fundamental difference between the past, present and future: the future, in contradistinction to the past and present, is "genuinely open" and harbors "real possibilities". These real possibilities are represented as different *branches*, i.e. different possible future histories, in a tree-like structure. The single trunk of this tree is the past; at the point on the trunk corresponding to the present, branching takes place; and the open future consists of (further branching) branches that develop independently of one another. The wider metaphysical context of this branching time framework is that of indeterminism and modal realism. The branching process is governed by probabilities that are understood as irreducibly modal single-case propensities, that measure the "graded powers" governing the transitions occurring during branching.

The metaphysical framework of genuine indeterminism, openness of the future and dispositional powers is in this way associated with a definite mathematical framework, namely that of a tree-like structure in which history branches into different possibilities at instants when indeterministic processes take place (in the context of relativity theory the picture has to take account of the absence of an absolute simultaneity relation and therefore becomes slightly more complicated; the branching then must happen at single space-time points rather than along simultaneity slices representing global instants – we shall not go into this in any detail, see Sects. 2.1 and 5.2 of Placek's paper).

The fundamental indeterminism that is needed to make sense of fundamentally dispositional powers and modal realism seems to tally well with quantum mechanics, and so Tomasz Placek, and Tomasz Placek together with Thomas Müller in other papers (see the references in Placek 2014), have set themselves the task of explicitly defining a mathematical structure of the kind just described, representing branching time, with appropriate "causal" probabilities defined on it, within the framework of quantum mechanics. The motivating idea behind this is that the applicability of such a formal causal framework to quantum mechanics will furnish a natural interpretation of quantum mechanics in terms of powers; and that the success of quantum mechanics will thus in turn support a powers ontology.

2 Quantum Theory, Branching Time and Branching Space-Time

The branching time conceptual framework from which Tomasz Placek's program starts presupposes a quite specific type of physical theory, namely a *space-time theory*, i.e. a theory defined on a space-time manifold. Such theories describe the world as basically consisting of space-time points, on which physical quantities are defined. In space-time theories the places where branching occurs can at least

be *located* in a natural way – on simultaneity hypersurfaces (three dimensional worlds-at-a-time) in a pre-relativistic setting, and in space-time points in a relativistic setting, respectively. The branching process itself is much more difficult to capture: the trunk representing the past becomes longer when the universe keeps on branching and this seems to require the notion of a shifting Now; but such a "motion" of the Now is notoriously difficult to reconcile with physics, see, e.g., Dieks (2012). It is not quite clear, however, whether the branching time theory defended by Tomasz Placek necessarily presupposes this "flow of time": one could think of a tenseless version of the theory in which all possible tree-structures, with all possible lengths of the trunks, are part of one "super-block universe". Past, present and future would in this case not be absolute notions, but would become relative to a tree with a given trunk length. It is not immediately clear how this would relate to modal realism (supposedly, the modalities should have to become relative as well). But since this point is not discussed in the paper commented upon here, we shall not go into this potential difficulty but rather focus on other aspects of the branching time project.

It is true that space-time theories are quite natural from the point of view of classical physics (and common sense, one may add), and also from the point of view of relativity theory. However, in quantum theory it is not at all self-evident that it is appropriate to think of the physics as defined against the background of a space-time manifold. The standard formalism of quantum mechanics is in terms of an abstract state space, Hilbert space, in which spacetime points play no basic physical role. Think of the discussions surrounding the notorious uncertainty relations: their standard interpretation is that in general it cannot be maintained that quantum objects possess definite positions – they are not placed in space the way classical objects are. It is true that in *relativistic quantum field theory* the attempt *is* made to formulate the theory mathematically against the background of a space-time manifold, but one should not be deceived by formal appearances. The fields of quantum field theory are smeared-out *operator* fields, and physical quantities like particle positions and momenta are not defined as functions on space-time points. So the status of space and time is here substantially different from that in classical theories (in the sense of non-quantum theories). In recent work in relativistic quantum field theory there are many attempts to explain the "emergence" of space and time, in the way we know them from classical physics and daily experience, as a kind of macroscopic phenomenon in the classical limiting case. The implication is that in the relevant quantum theories space and time do not possess a basic status from a fundamental theoretical point of view.

So, if one comes from the side of quantum physics, it is not obvious that anything like a branching time or branching space-time framework is appropriate. However, it might be objected that this conclusion takes a more or less standard interpretation of quantum mechanics for granted, and that other interpretations have been proposed that are more sympathetic towards classical space and time. In particular, in the 1950s David Bohm (1952) developed a "hidden-variables" theory according to which particles do possess determinate spatial positions, just as classical particles

(these well-defined positions do not occur in standard quantum mechanics and in this sense are "hidden" from view – hence the terminology). Bohm's theory succeeds in reproducing the same empirical predictions as standard non-relativistic quantum theory, so it is empirically equivalent. But empirical equivalence does not automatically entail theoretical underdetermination – a well known theme from modern philosophy of science. In fact, even in cases of empirical equivalence there may be weighty *empirical* reasons to prefer one theory over another (Laudan and Leplin 1991; Acuña and Dieks 2014). Bohm's theory is a case in point. Its structure flies in the face of what we have learned from the history of physics: most importantly, it incorporates a classical-like action at a distance, which causes problems for relativistic generalizations, and contains a link between ensemble densities, probabilities and interaction potentials that has never been seen before in physical theory. These are features that are quite out of line with the conceptual structure of modern physics: there is no independent *empirical* support that the physical world is anything like that. This is not to deny, of course, that *metaphysical* predilections could make such a theory attractive nevertheless.

In fact, the space-time structure of Bohm's theory seems to provide an appropriate arena for the introduction of branching. However, Bohm's interpretation of quantum mechanics is not the one considered by Placek, perhaps because of the difficulties this theory is confronted with in the face of relativity theory (see the beginning of Sect. 2.1 of Placek's paper); the status of probability in Bohm's theory is likely to be an even more important obstacle, see Sect. 3 below. Instead, Placek turns to the GRW theory of spontaneous localization (Placek 2014, Sect. 2), following Dorato and Esfeld (2010).

It should be noted that this GRW theory is not an *interpretation* of quantum mechanics, at least not in the sense that it is a scheme that yields the same empirical predictions as the standard formalism. In fact, the GRW theory introduces two new constants of nature: one, λ, governing the time rate at which indeterministic localizations take place, and the other, σ, characterizing the width of the quantum mechanical wave packets to which these localization processes lead. These two new constants specify two *new* processes that do not occur in standard quantum mechanics. The associated change in the formal structure of the theory has the consequence that the GRW predictions are not exactly the same as those of ordinary quantum mechanics. However, the values of λ and σ have been chosen such that the differences from standard quantum mechanics are hard to verify experimentally; they cannot (yet) be detected with present-day experimental techniques.

The general idea of the GRW theory is that the wave function of a many particles system develops according to the standard quantum mechanical evolution equation, the Schrödinger equation, but that from time to time, in a random process with time rate λ, the wave function *collapses* to a set of narrow wave packets of size σ. These narrow wave packets seem plausible candidates for an interpretation in terms of things that are more or less localized in space (something needed in the branching time account). In the original GRW theory the narrow wave packets were interpreted as representing small spatial areas with a non-vanishing mass density, which persist over time. However, it is not easy to generalize this "material",

particle-like interpretation into a natural relativistic version, and this is one of the reasons why Placek opts for another interpretation: an ontology in terms of "flashes", chancy events occurring at the places where the wave packets (more or less) localize.

It is not completely clear what the flashes are flashes of; but in order to ensure, in the classical and macroscopic limit, agreement with experience it appears that we have to think of "particles-at-an-instant", instantaneous stages of what ordinarily would be the history of a particle. However, it is important to realize that in this flash ontology there *are* no particles with histories: it is only the *illusion* of continuous particle paths that is produced in the limiting case in which the localizations become very frequent (this happens when the N of the "N particles wave function" becomes very large, of macroscopic order). So the situation is somewhat analogous to that of a movie, in which a rapid succession of frames can evoke the impression of continuous motion, although in reality there is only a discontinuous series of motion-stages. In the GRW theory there is a discontinuous series of flashes, with nothing in between, which creates the impression of particle motion. The absence of real particles with continuous paths is essential here: it is exactly this feature that makes the flash ontology amenable to relativistic generalization.

The flashes provide us with the indeterministic space-time events in which we can imagine branching to take place, governed by "graded causal powers", and this make the GRW theory attractive from the point of view of Tomasz Placek's program. As he interprets GRW, (Placek 2014, Sect. 3): "... a power anchored in a configuration of flashes dictates the degrees of possibility of later configurations of flashes."

3 A Critique

We mentioned, in connection with Bohm's theory, that even in the case of empirically equivalent theories there may be strong reasons for preferring one theory over another. There are the usual arguments for this conclusion based on theoretical virtues like simplicity, elegance and explanatory power – however, one may object that such arguments are not fully objective and that the virtues in question are at least partly pragmatic in character. But stronger criteria are available, which relate more directly to the overriding aim of empirical science to take all empirical evidence fully into account. Since Laudan and Leplin's seminal paper (Laudan and Leplin 1991) this has become a generally accepted principle of confirmation theory, see also Acuña and Dieks (2014). Most importantly, one can look at how theories achieve with respect to general properties of the physical world that can be gathered from empirical background knowledge, and at how theories fit into more general empirically confirmed theoretical schemes. In this way one can argue that in spite of the empirical equivalence, Bohm's theory displays properties that are less well empirically confirmed than the structural properties of standard quantum mechanics.

This pattern of reasoning applies a fortiori to the GRW theory. The GRW theory is *not* empirically equivalent to standard quantum mechanics, but achieves

a for-practical-purposes indiscernibility by postulating an upper bound of the values of the discrepancies. The existence of the spontaneous collapse mechanism, which is responsible for the differences between quantum mechanics and the GRW theory, possesses no empirical support. This newly postulated physical mechanism breaks the symmetry of quantum theory by introducing a privileged status for the physical quantity "position" (the collapses are to approximate position eigenstates). But it is exactly the symmetrical Hilbert space formalism that has led to great success, also on the empirical level, in generalizing quantum mechanics to the standard model of elementary particles and to the quantum field theories that are now offering prospects of incorporating gravity (e.g., via the AdS/CFT duality).

Moreover, empirical evidence is accumulating that even when we come closer and closer to the macroscopic level, quantum mechanical superpositions are able to keep on maintaining themselves (in so-called macroscopic Schrödinger cat states) – whereas it is the very core idea of the GRW theory to introduce a mechanism that breaks such superpositions. So although an immediate empirical refutation of the GRW theory seems not yet feasible with the help of present-day laboratory techniques, there are good scientific, inductive empirical reasons for being suspicious of the theory.

Seen from this vantage point, the situation appears rather clear: the attempt to model quantum mechanics after the example of branching space-time theories boils down to imposing a metaphysical system, deriving from everyday experience and untutored common sense, on a sophisticated modern physical theory – even to the extent that scientifically unwarranted changes to this theory are embraced if these changes fit the presupposed metaphysics. This is exactly the kind of manoeuver criticized by Mach when he forcefully objected against the interpretation of classical mechanics in terms of anthropomorphic, normative, "animistic" concepts.

However, we have until now neglected a weighty argument that might be used to back up Placek's enterprise. In fact, theories like Bohm's quantum mechanics and the GRW theory have not been cooked up by metaphysicians: they have been proposed by *physicists* in an attempt to avoid a problem that is internal to quantum theory, namely the infamous *measurement problem*. The measurement problem, in a nutshell, is that quantum mechanics speaks about the probabilities of finding certain results *in a measurement*, but that it is not unambiguously specified what exactly a measurement *is*. Is it necessary for a measurement that a macroscopic measuring device is interacting with an object system, or perhaps even that a conscious observer is taking part in the process? It would certainly be a serious mistake to think that the latter requirement is part of standard quantum mechanics (loosely speaking: quantum mechanics in the Copenhagen tradition), but concerning the role of a measuring device and whether or not this device should be macroscopic the situation is certainly unclear. Given this ambiguity, it is also unclear what the indeterministic events are the quantum probabilities refer to. This lack of clarity does not hamper day-to-day physical practice, because for this it is sufficient to know that the theoretical probabilities materialize in frequencies and expectation values in laboratory experiments. But from a foundational viewpoint one would like to know more: what are *in general*, even in situations in which no laboratory experiments are

being performed, the chancy events to which the quantum probabilities attach? It is to answer *this* question that various "interpretations" of quantum mechanics have been proposed. The interpretations of quantum mechanics are attempts to solve the measurement problem: they specify what the indeterministic occurrences are that quantum mechanics is about, even in cases in which there are no measurements in the sense of human interventions.

Bohm's interpretation is one example of such an interpretation: it says *that there are no chancy events in Nature at all*, and that the quantum mechanical probabilities refer to our *lack of knowledge* about the initial conditions of the (deterministic) evolution of particles (this makes Bohm's interpretation unattractive for the branching space-time program). The GRW theory, as we have seen, is not really an interpretation of quantum mechanics but rather a rival theory; in the version that concerns us here it specifies *flashes* as indeterministic events to which probabilities refer. But there are many more interpretations of quantum mechanics. Examples are the many-worlds interpretation, modal interpretations, the consistent histories interpretation, several "relational" interpretations, and proposals based on "decoherence". These are all genuine interpretations of quantum mechanics, in the sense that they are empirically equivalent to the standard theory; they differ, in some cases in rather subtle ways, in how they specify – in the most general situations – the events to which the quantum probabilities pertain. What these interpretations have in common, except for Bohm's interpretation, is that they do not assign any a priori privileged position to spatial localization but respect the theoretical structure of quantum mechanics according to which all physical quantities basically occur in a symmetric fashion.

Therefore, although the GRW theory was originally proposed as one possible response to the measurement problem, it is certainly not the only possible response, nor the most natural one from the quantum point of view. The situation surrounding the measurement problem remains controversial, but it seems safe to say that there is not much on the physical and empirical side that makes the GRW scheme plausible. A preferred status for the GRW theory must be argued for on grounds coming from outside physics, even if we take the measurement problem into account.

4 Metaphysics and Mathematics

It would therefore be far-fetched to interpret the possibility of an interpretation of the GRW theory in terms of a "branching time cum causal powers ontology" as support from the side of quantum mechanics for such an ontology. A more natural conclusion to draw would rather be that apparently quite a drastic adaptation of quantum mechanics is needed to make quantum mechanics compatible with branching time. Actually, the situation is worse for the defender of branching time plus causal powers. Indeed, it is unclear in what way the causal powers ontology receives support even if we accept the GRW theory and Tomasz Placek's analysis of it.

First, it remains obscure whether compatibility between an ontology of powers and GRW has been really demonstrated in the analysis. In Sect. 3 of his paper Placek writes: "Powers are anchored in concrete objects (particles, fields). Powers are modal, and require the so-called real possibilities (aka historic possibilities). An object's power rules (...) what is possible, and what is not for this object. Finally, an inherent part of powers is the gradation of possibilities: in our case, a power anchored in a configuration of flashes dictates the degrees of possibility of later configurations of flashes". And a bit later, in Sect. 3.1, he continues in the same vein: "..., since modal powers need to be anchored in objects, one needs to develop a modally-thick notion of objects, the notion that conceives of objects as having alternative future developments, some of which become actualized in due course, and some of which do not". But as Placek himself acknowledges in his Conclusion, the GRW scheme does not contain objects in any ordinary sense, and therefore appears to lack the bearers of modal powers that were declared to be necessary for the success of the modal project: there are neither particles nor fields in the GRW flash world. A way out that seems possible at first sight, and that is hinted at in the first of the above quotes, is to consider *a configuration of flashes* as itself an object. But this suggestion leads to several immediate problems. First, in the absence of a notion of absolute simultaneity it is unclear how to group flashes together to form one object: the notion of a "configuration of flashes" is ill-defined in the relativistic context. Second, even if we cut this knot in some way and define flash-objects via some grouping-together recipe, the dilemma remains that the objects that are formed this way do not possess continuous histories. Between one configuration of flashes and another one there is literally *nothing*, so that one is at a loss to understand how the power of the earlier configuration makes itself felt at the position of the later one. Since one of the claimed advantages of the introduction of powers (at least as I understand it) is the possibility of giving an *explanation* for the occurrence of chance events, this seems an important objection.

But there is a second and more general qualm concerning the strategy of the causal probability project. The real work in Tomasz Placek's paper is done in Sect. 5, "Probability next" (since modality is assumed to come first). But this core work of the paper is of a purely *mathematical* nature. What is demonstrated is that the GRW localization probabilities can be embedded in a mathematical framework in which probabilities can be consistently assigned to alternate histories; in other words, a mathematically respectable probability space is defined. Now it is certainly true that the explicit construction and description of this probability space is by no means trivial, and it is very nice to see the job done. But it seems to me that the solvability of this problem was clear from the outset: it only depends on the probabilistic *consistency* of the GRW theory. The GRW theory was meant to generate different possible future developments, with different probabilities, and if this intention was implemented in a mathematically consistent way it *should* be possible to define a probability space whose elements are possible histories, with consistent probabilities attached to them. If it had turned out, as a result of Placek's analysis, that such a probability space were not possible, then this would have signified the inconsistency and therefore untenability of the GRW theory. Exactly

the same remark applies to all dynamic stochastic theories, and to the various interpretations of quantum mechanics that we mentioned above. To the extent that these theoretical schemes define mutually exclusive histories, there should be a probability space containing these histories, on pain of inconsistency.

Now, the important point is that this simple consideration is completely independent of the *interpretation* of the alternate histories or the probabilities attached to them. It does not matter whether we are realists about modalities and perhaps also about possible worlds, or whether we are strict nominalists who think that all talk about modalities is just verbal, without ontological reference. Our theory of probability had better be consistent anyway!

If this is right, then how can the construction of a consistent probability space be construed as any support at all for an ontology of causal powers? If the probability space had proved non-existent, *all* readings of the probabilistic theory would have fallen, together with the theory itself. Nothing in particular follows about the interpretation of probabilities in terms of causal powers, or any other interpretation of probability for that matter.

5 Conclusion

Tomasz Placek has made an interesting attempt to bridge the gap between physics and metaphysics, by linking the doctrines of branching time and modal realism to quantum theory. The idea of different branches of history has certainly an intuitive plausibility, and the same can be said of the conception that chance events do not just happen, but are governed and can be explained by the action of causal powers. Given the importance of indeterminism in these doctrines, it is natural to turn to quantum mechanics for support. However, I do not think that Placek has succeeded in showing that quantum mechanics gives additional plausibility to either branching time or a powers ontology. The arguments for these metaphysical doctrines remain purely philosophical and quantum physics is at best neutral, and at worst difficult to reconcile with them.

References

Acuña, P., and D. Dieks. 2014. Another look at empirical equivalence and underdetermination of theory choice. *European Journal for Philosophy of Science*, doi:10.1007/s13194-013-0080-3.

Bird, A. 2010. *Nature's metaphysics: Laws and properties*. Oxford: Oxford University Press.

Bohm, D. 1952. A suggested interpretation of the quantum theory in terms of 'hidden variables', I. *Physical Review* 85: 166–179; II. *Physical Review* 85: 180–193.

Dieks, D. 2012. The physics and metaphysics of time. *European Journal of Analytic Philosophy* 8: 103–120.

Dorato, M., and M. Esfeld. 2010. GRW as an ontology of dispositions. *Studies in History and Philosophy of Modern Physics* 38: 371–389.

Harré, R., and E.H. Madden. 1975. *Causal powers: A theory of natural necessity*. Oxford: Blackwell.

Laudan, L., and J. Leplin. 1991. Empirical equivalence and underdetermination. *The Journal of Philosophy* 88: 449–472.

Lewis, D. 1986. Causation. In *Philosophical papers*, vol. 2, 159–213. Oxford: Oxford University Press.

Placek, T. 2014. Causal probabilities in the GRW quantum mechanics. In *New directions in the philosophy of science*, this volume, ed. M.C. Galavotti, et al. Cham: Springer.

Part V
History of the Philosophy of Science

Where Would We Be Without Counterfactuals?

Huw Price

1 Remembering Russell

One hundred years ago,[1] on 4 November 1912, Bertrand Russell delivered the Inaugural Address to the Aristotelian Society's 34th session – it was Russell's second year as President. His lecture was entitled "On the Notion of Cause", and the Society's *Proceedings* inform us that a discussion followed in which a number of members took part. Outside the Aristotelian Society, the discussion continues to this day. Russell's paper remains both influential and controversial. It is widely known as the source of one of the most famous lines in twentieth century philosophy: "The law of causation", Russell declares, "Like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." (Russell 1913, p. 1)

As it happens, "On the Notion of Cause" is the one piece of Russell's entire philosophical output with which my own work connects directly – certainly the only one where I make any claim to advancing the matters under discussion. So I want to take advantage of this happy coincidence to use my own inaugural lecture, to the Cambridge professorship that now bears Russell's name, to celebrate the centenary of this famous paper; and to say something about what its conclusions look like, at least from my perspective, with the benefit of a century's hindsight.

[1]This piece was written for delivery on 1 November 2012, as my inaugural lecture as the Bertrand Russell Professor of Philosophy, Cambridge. I am very grateful to Maria Carla Galavotti and the organisers of "New Directions in Philosophy of Science" for the invitation that enabled me to give the lecture as a dress rehearsal in Bertinoro on 20 October 2012, and to the audience on that occasion for many helpful comments and questions.

H. Price (✉)
Faculty of Philosophy, University of Cambridge, Sidgwick Avenue, CB3 9DA Cambridge, UK
e-mail: hp331@cam.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_42,

It is a story with a lot of Cambridge connections. Indeed, I'm not the first occupant of what is now the Bertrand Russell Chair to mention Russell's paper in an inaugural lecture. Anscombe does so in her lecture "Causality and Determination", from 1971:

> Russell wrote of the notion of cause, or at any rate of the 'law of causation' (and he seemed to feel the same way about 'cause' itself), that, like the British monarchy, it had been allowed to survive because it had been erroneously thought to do no harm. In a destructive essay of great brilliance he cast doubt on the notion of necessity involved, unless it is explained in terms of universality, and he argued that upon examination the concepts of determination and of invariable succession of like objects upon like turn out to be empty: they do not differentiate between any conceivable course of things and any other. (Anscombe 1971, p. 135)

This brings Anscombe to her own concern, namely, indeterministic conceptions of causality, and it gives her a reason to take Russell to task: "Thus Russell too assumes that necessity or universality is what is in question, and it never occurs to him that there may be any other conception of causality." (*ibid.*)

My concerns are different from Anscombe's, but I will follow her in one respect, interpreting Russell's target broadly rather than narrowly. Anscombe takes it to include "cause", as well as "the law of causation". I'm going to take it to include a broad class of counterfactual judgements – judgements of the form "If A had not happened, the B would not have happened." I don't suggest that Russell himself took his thesis to extend this far; but the history of the subject since Russell's time has shown that such counterfactuals are closely connected not only to causation in general, but to some of the specific issues about causation at the core of Russell's paper.

One of Russell's key arguments concerns the *time-asymmetry* of causation – the fact that effects are supposed to occur *later* than their causes (or at least *not earlier* than their causes). Russell says that there is nothing to ground such a difference between past and future in fundamental physics. More recent writers have noted that there is a similar time-asymmetry in counterfactual reasoning. If the lights had failed at the beginning of this lecture, the proceedings since that moment might well have been different – I would not have progressed this far, for example – but events before the failure of the lights would have just the same; or at least, so our intuitions tell us. David Lewis (1979) calls this the asymmetry of counterfactual dependence, and proposes to use it to *explain* the asymmetry of causation, that Russell had taken to be missing in fundamental physics.

A good question at this point is where the time-asymmetry of counterfactual dependence comes from, if Russell is right, and there isn't anything suitable in fundamental physics? I'll come back to this. For the moment, I've mentioned it to make the point that we can't sensibly discuss the issues raised by Russell's paper without broadening the scope in this way, to include counterfactuals.

But broadening the scope raises the stakes. Russell seems to be suggesting that abandoning talk of causality would be practical, as well as desirable, much as abandoning the monarchy might be: some planning, a few days of mild confusion,

and it would all be done. Indeed, he thought that modern physicists, vanguards of the revolution, had already made the change:

> All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never occurs. (*ibid.*)

Russell then singles out for criticism his former teacher, James Ward, who was actually the very first holder of this chair – its inaugural inauguree, so to speak. Russell says that Ward makes the fact that the advanced sciences don't mention causation "a ground for complaint against physics: the business of science, he apparently thinks, should be the discovery of causes, yet physics never even seeks them." Russell replies: "To me it seems that philosophy ought not to assume such legislative functions, and that the reason why physics has ceased to look for causes is that, in fact, there are no such things." (*ibid.*) (After that, we get the famous line about monarchy.)

Now, it is debatable whether Russell was right even about physics – or whether, even if he was right about the advanced physics of his own time, what he saw was actually a general feature of future (presumably even more advanced) physics. Forty years ago, Patrick Suppes argued that Russell's claim was not true of the physics of the 1960s.

> Contrary to the days when Russell wrote this essay, the words 'causality' and 'cause' are commonly and widely used by physicists in their most recent work. There is scarcely an issue of *Physical Review* that does not contain at least one article using either 'cause' or 'causality' in its title. (Suppes 1970, pp. 5–6)

But even if Russell is right about physics, the idea that we could dispense with talk of causality much as we might dispense with the monarchy still seems wildly unrealistic.

The point is even more obvious when we notice the way in which counterfactuals are liable to be swept up in the same net. For my text at this point I take some wise words I once heard attributed to the distinguished American philosopher, Jerry Fodor. "Why is real estate in Manhattan so damned expensive?", Fodor asks – "You're paying for all those counterfactuals!" As usual from Fodor, it's a pithy little piece of philosophy, served with a generous *amuse bouche*. And he's right, to a considerable extent, obviously.

But imagine the consequences if Fodor's news and Russell's news leak out at the same time. The value of your apartment depends on its counterfactuals, but Bertrand Russell says there ain't no counterfactuals! So the whole Manhattan market is a gigantic Ponzi scheme, built not on sand, not even on paper, but literally on *nothing!* It's unthinkable – and literally so, perhaps, if counterfactual thought is to be imagined groundless, and yet we need a counterfactual to ask the question, as in my title.

At this point we might defend Russell by suggesting that his real target is much more modest. His view about the word "cause", as he puts it, is that it "is so inextricably bound up with misleading associations as to make its complete

extrusion from the *philosophical* vocabulary desirable" (*ibid.,* my emphasis), and that there are no causes at a fundamental level in physics. But this is compatible, perhaps, with recognising a non-fundamental role for causal vocabulary (and counterfactuals too), in non-fundamental parts of physics and in everyday life. Provided such a role could be found, then we might save the real estate market, indeed the finance system itself – for what is money except a convenient form of exchangeable counterfactuals? – without treating talk of causation and counterfactuals as tracking some deep, God-given feature of the furniture of reality. (After all, remember the first three laws of real estate – "Location, location, location". They surely survive the discovery that there is no such thing as absolute space!)

So we can defend Russell by taking him, like Hume before him perhaps, to be proposing only an armchair revolution. He wants to banish "causation" from the kind of serious conversation that goes on around the Great Court at Trinity College (and similarly venues elsewhere, if such there be), without banishing it from the streets. (Where contemporary physics falls under this regime would be a matter for debate. We would need to figure out what physicists *mean* when they talk of causation, and that might be a lengthy investigation.)

But even this modest armchair revolution seems to be undone by a famous paper by Nancy Cartwright, her "Causal Laws and Effective Strategies" (Cartwright 1979). Cartwright begins with Russell's distinction between symmetric laws of association – the kind of Humean regularities Russell thinks that modern physics actually offers us – and the asymmetric causal laws Russell thinks we need to dispense with (at least in the armchair). She notes that Russell argues for two conclusions: in her words, (i) that "laws of association are all the laws there are", and (ii) "that causal principles cannot be derived from the causally symmetric laws of association". She goes on to argue "in support of Russell's second claim, but against the first." (1979, p. 419) That is, she agrees that "[c]ausal principles cannot be reduced to laws of association", but maintains that "they cannot be done away with." (*ibid.*)

Cartwright's argument is that causal laws are needed to ground an important distinction between effective and ineffective strategies. She illustrates this distinction with some examples, one of them a letter she tells us she received from an insurance company, making the following claim:

> It simply wouldn't be true to say, 'Nancy L.D. Cartwright … if you own a TIAA life insurance policy, you'll live longer.' But it is a fact, nonetheless, that persons insured by TIAA do enjoy longer lifetimes, on the average, than persons insured by commercial insurance companies that serve the general public. (1979, p. 420)

Cartwright argues that the objective fact reported in this letter – viz., that buying life insurance from this company would *not* be an effective strategy for living longer – depends on *causal* rather than merely *probabilistic* facts about the world. But, she argues, the "objectivity of strategies requires the objectivity of causal laws". In other words, she concludes,

> causal laws cannot be done away with, for they are needed to ground the distinction between effective strategies and ineffective ones. … [T]he difference between the two depends on the causal laws of our universe, and on nothing weaker. (*ibid.*)

Commenting on Cartwright's argument in a recent survey paper, Hartry Field concludes:

> This makes a compelling case against Russell's view that we should do without causal notions. But Cartwright herself draws a much stronger conclusion, a kind of *causal hyper-realism*, according to which there are causal facts *that outrun the totality of 'noncausal facts'* (i.e. the facts that could be expressible in some language without using causal terminology). Indeed, her claim isn't simply that there is no reasonable way to explicitly define causation in noncausal terms; it seems to be that causal claims don't even supervene on the noncausal facts. Among the 'noncausal facts' she includes the basic laws of physics – e.g. Newton's law that an object accelerates in direct proportion to the force impressed on it and in inverse proportion to its mass. She holds that the causal fact that a force on an object *makes* the object go faster is not reducible to Newton's law, nor to other noncausal facts either, such as the equations of energy flow from the sources of fields to the fields themselves to the accelerating objects. (Such equations are just further parts of fundamental physics, which she regards as 'laws of association' rather than as causal.) Rather, the claim that a force on an object makes the object go faster states a further truth about the world that physics leaves out. Evidently there is some sort of causal fluid that is not taken account of in the equations of physics; just how it is that we are supposed to have access to its properties I am not sure. (Field 2003, Sect. 2)

Field finds this hyper-realism unpalatable, but recognises the importance of Cartwright's challenge:

> But despite the implausibility of the hyper-realist picture, we have a problem to solve: the problem of reconciling Cartwright's points about the need of causation in a theory of effective strategy with Russell's points about the limited role of causation in physics. *This is probably the central problem in the metaphysics of causation.* (*ibid.,* Sect. 2, emphasis added)

For my part, I agree with Field, except that I think it would be better to say that it is one of *two* central problems, the other being Russell's issue of the time-asymmetry of causation. Accounting for that is a further difficulty for the hyper-realist view, but it is a problem for all views, including views which want to regard causation as something relatively non-fundamental.[2]

2 Russell on Time-Asymmetry

Russell's own treatment of the problem of the apparent time-asymmetry of causation is important more because he sees that there is a problem than because he provides any satisfactory solution, or dissolution. Let's have a look at the central passage. Russell puts his cards on the table straightaway:

> We all regard the past as determined simply by the fact that it has happened; but for the accident that memory works backward and not forward, we should regard the future as equally determined by the fact that it will happen. (Russell 1913, pp. 20–21)

[2]This problem has also been much discussed in recent literature, especially by way of criticism of David Lewis's attempt, mentioned above, to explain a corresponding asymmetry of counterfactual dependence, to which that of causation might then be reduced. See, e.g., Price and Weslake (2009) and references therein.

In other words, so Russell is claiming, there's no real distinction between past and future, but an accident about us – the fact that memory works backwards not forwards – tricks us into thinking that there is. He then counters his own claim with a plausible objection: "'But,' we are told, 'you cannot alter the past, while you can to some extent alter the future.'" (*ibid.*)

In reply to this Russell makes two points. The first that while it is true that "you cannot make the past other than it was", this is just a matter of logic, and the same is true with respect to the future. The second – acknowledging, I think, that the first has not got to the heart of the matter – introduces what turns out be an important link between what we think we can influence and what we can *know*:

> [I]f you happen to know the future – *e.g.,* in the case of a forthcoming eclipse – it is just as useless to wish it different as to wish the past different. (*ibid.*)

Russell's self-scripted interlocutor now does a good job of getting things back on track:

> 'But,' it will be rejoined, 'our wishes can *cause* the future, sometimes, to be different from what it would be if they did not exist, and they can have no such effect upon the past.' (*ibid.*)

(Even though we often don't know the past, if would have been helpful to add!)

At this point, I think, Russell loses his grip on the force of his opponent's argument. What he says is just this:

> This, again, is a mere tautology. An effect being *defined* as something subsequent to its cause, obviously we can have no *effect* upon the past. (*ibid.*)

But we can just give Russell the terms "cause" and "effect", defined in this way, and press the original objection as the question as to why it never makes sense to act for (or "wish for") ends which – while not *effects* of our actions (being ruled out as such by this definition, because they lie in the past) – are nevertheless desirable, from our point of view. Why is it useless to do something now to ensure that my ticket was the winning ticket in a lottery drawn yesterday, for example (if I don't yet know that it is not)? Call this the problem of the *time-asymmetry of deliberation*. Russell doesn't get this problem in focus, I think. You can't avoid the question by saying that an end for which we act is *by definition* an effect of our action, because then you'd have two definitions of "effect" in play, and you'd need to explain why they line up – and that's just the original problem.

There then follows an interesting little passage I won't go into in detail, in which Russell is in effect denying what we now call the asymmetry of counterfactual dependence:

> But that does not mean that the past would not have been different if our present wishes had been different. Obviously, our present wishes are conditioned by the past, and therefore could not have been different unless the past had been different; therefore, if our present wishes were different, the past would be different. Of course, the past cannot be different from what it was, but no more can our present wishes be different from what they are; this again is merely the law of contradiction. (*ibid.*)

And Russell then sums up:

> The facts seem to be merely (1) that wishing generally depends upon ignorance, and is therefore commoner in regard to the future than in regard to the past, (2) that where a wish concerns the future, it and its realization very often form a 'practically independent system,' *i.e.,* many wishes regarding the future are realized. But there seems no doubt that the main difference in our feelings arises from the fact that the past but not the future can be known by memory. (*ibid.*)

I want to emphasise three points:

1. Russell is trying to explain the apparent time-asymmetry of causal dependence as a product of a difference in us, rather than a fundamental difference in reality. (As I'll explain, I think he's right about that.)
2. The difference he picks concerns memory, but it's doubtful whether that can do the trick – it doesn't draw a clean enough cut between past and future, because we remember rather little of the past, and know some of the future by other means.
3. He already has on the table the idea that the business of realization of our wishes, as he quaintly puts it, is important in explaining the illusion of asymmetry – though, hampered by the thought that it is all about memory, he doesn't get very far.

And that positions us for the next step in this story, which belongs to perhaps the greatest of all these early twentieth century Cambridge giants, Frank Ramsey.

3 Ramsey's Special Agent

The step in question occurs in Ramsey's very late paper, "General Propositions and Causality", a rough 30-page manuscript dated September 1929 (just 4 months before Ramsey's tragically early death, at the age of 26). About two-thirds of the way through, Ramsey turns to the issues of the difference between past and future and the direction of causality. Then, in a few short paragraphs – little more than two handwritten pages – he dissects his way to core of the problem, and shows us what he takes to lie at its heart. In my view, this is one of the most insightful passages not merely in twentieth century philosophy, but in all of philosophy. Indeed, given the centrality of the issues at stake here – causality, the direction of time, and the location, in these respects, of the cut between what properly belongs in the world, and what it some sense "stems from us" – I think it has some claim to be a significant landmark in human thought as a whole.

As usual, however, Ramsey is here "too brisk for most philosophers", as Hugh Mellor (1988, p. 254) said of a different piece of Ramsey, in his own inaugural lecture to this chair. And unusually, I think, he doesn't quite get everything right – what survives of his view is just the back of the envelope version, after all. But he does give us an answer to both the puzzles we have on the table. I'll give you the core of the relevant passage, highlighting what I take to be some key remarks, and then explain these points in more detail.

> It is, it seems, a fundamental fact that the future is due to the present, or, more mildly, is affected by the present, but the past is not. What does this mean? It is not clear and, if we try to make it clear, it turns into nonsense or a definition …
>
> [W]e think there is some difference between before and after at which we are getting; but what can it be? There are differences between the laws deriving effect from cause and those deriving cause from effect; but can they really be what we mean? No; for they are found *a posteriori,* but what we mean is *a priori.* [The Second Law of Thermodynamics is a posteriori; what is peculiar is that it seems to result merely from absence of law (i.e. chance), but there might be a law of shuffling.]
>
> What then do we believe about the future that we do not believe about the past; the past, we think, is settled; if this means more than that it is past, it might mean that it is settled *for us,* that nothing now could change our opinion of it, that any present event is irrelevant to the probability for us of any past event. But that is plainly untrue. What is true is this, that **any possible present ~~action~~**[3] **volition of ours** is (for us) irrelevant to any past event. To another (or to ourselves in the future) it can serve as a sign of the past, but **to us now what we do affects only the probability of the future.**
>
> This seems to me the root of the matter; that I cannot affect the past, is a way of saying something quite clearly true about my degrees of belief. Again **from the situation when we are deliberating seems to me to arise the general difference of cause and effect.** We are then engaged not on disinterested knowledge or classification (to which this difference is utterly foreign), but on tracing the different consequences of our possible actions, **which we naturally do in sequence forward in time,** proceeding from cause to effect not from effect to cause. We can produce A or A′ which produces B or B′ which etc. … ; the probabilities of A, B are mutually dependent, but *we* come to A first from our present volition. … **In a sense my present action is an ultimate and the only ultimate contingency.** (Ramsey 1929, pp. 145–146, emphasis in bold added)

Let us explore some of the ideas we find in this extraordinary passage.

3.1 Agency Is the Key to Understanding Causation

Ramsey thinks that causation – especially the asymmetry of causation, the difference between cause and effect – needs to be understood by thinking about "the situation when we are deliberating." This idea turns up in later writers, including at least three with Cambridge connections. The first of these was Douglas Gasking, a student in Cambridge the 1930s, and later a lecturer at the University of Melbourne, who defends the idea in a wonderful paper in *Mind* in 1955 (Gasking 1955); the second was G.H. von Wright, briefly Wittgenstein's successor in this chair, who defended such a view in the 1970s (von Wright 1973, 1975). In Oxford, Collingwood (1940) had proposed it in the 1930s. But so far as I know, none of these writers recognised that Ramsey was on to the idea, nor gets anywhere close to the insights that connect Ramsey's version of the view to the issues raised by Russell's paper.

[3]This deletion is present in the manuscript of Ramsey's paper.

Later, in the early 1990s, I also defended a version of this agency view, both in joint work with my Australian colleague Peter Menzies (Menzies and Price 1993) and elsewhere (Price 1991, 1993). This time Ramsey does get a mention, as do issues such as the time-asymmetry of causation, and Cartwright's challenge to Russell. More recently still, the idea that causation needs to be understood in terms of "manipulation", as people now say, has been very prominent over the past decade or so, thanks to the work of the philosopher Jim Woodward (2003) and the computer scientist Judea Pearl (2000), amongst others. But Ramsey was the first, so far as I know.

3.2 *Probability Looks* Different *from an Agent's Point of View*

Ramsey's key insight is that agents are epistemically "special" – probability judgements are properly *different* from an agent's first-person point of view, than from a third-person point of view (even that of the same agent at other times). Ramsey puts this initially as the claim that our own present volitions are, for us, probabilistically independent of "any past event". He doesn't actually tell us why this is so, though he does give us a hint at the end: "my present action is an ultimate and the only ultimate contingency".

What he has in mind, I think, is a generalisation of what is now a familiar point about how knowledge and free choice conflict. It doesn't make sense to take yourself to be choosing between options one of which you know to obtain. (Remember Russell's example of the eclipse, known in advance.) The point emerges in the oddity of saying something like, "I know I'll vote for Obama, but I'm still making up my mind." You take away with one hand the authority you bestow with the other, much as in Moore's paradox, when you say "P, but I don't believe that P."

Ramsey generalises this point from knowledge to credence, so that it becomes the claim that an agent cannot take herself to have evidence about what she is going to do, *as she deliberates*. ("Deliberation screens prediction", as Rabinowicz (2002) puts it.) From the agent's point of view, her contemplated action must be regarded as probabilistically independent of anything she does she know at present – even if *other people* (or she herself at other times) could legitimately take something of that kind as evidence about her choice, or vice versa.

In my view, this is the key to solving Field's puzzle. The differences Cartwright rightly points to between the probabilities we get from laws of association, on the one hand, and the probabilities associated with what we take to be causal dependencies, on the other, are precisely the differences induced by the specialness of agency. Cartwright is right in thinking that we can't explain effective strategy in terms of the former; but wrong to think that we need causal laws to give us the latter. On the contrary, as I think Ramsey first saw, it's the other way round: the specialness of the agent's perspective grounds our talk of causation.

3.3 The Time-Asymmetry of Causation

This leaves the problem of the time-asymmetry of causation, and here I think Ramsey is at least looking in the right direction, saying it depends "on tracing the different consequences of our possible actions, *which we naturally do in sequence forward in time*." The first comment we need to make here is that "consequences" must mean something like "probabilistic dependencies", if Ramsey is not to be accused of slipping in talk of causation at this point. Second, since Ramsey has already appealed to the idea that there are no such dependencies between our present actions and events in the past, the qualification "naturally" is in a sense unneeded: if Ramsey is right, there are no dependencies *except* "forward in time." (Imagine turtles hatched on an East-facing beach. Is it any surprise that their journey takes them asymmetrically to the East? No, for there are no routes to the West, from their point of view.)

Still, the term "naturally" is useful in another sense. It reminds us that we are appealing to a feature of our natures – a universal feature for us, albeit perhaps a contingent one. As structures in spacetime, we human agents all share a common temporal orientation. Imagine depicting our deliberative lives on a spacetime map, with a little arrow connecting each instance of deliberation to its associated action (where there is one). For us, all those little arrows point in the same direction – from *past* to *future*, as we would normally put it.

Or, to go back to the turtles, think of yourself as a beach, and of your own plans and deliberations as the turtles that hatch on that beach. What's true is not only that all plans hatch in the same direction on each beach individually, but also that all the beaches we know of face in the same direction. It is an interesting question whether there could be creatures elsewhere whose beaches face in the other direction – whose plans hatch from future to past, by our lights. The answer is probably yes, in my view, so long as they are far enough away in spacetime to live in a region with low entropy in (what we call) the future, rather than (what we call) the past, but that's a long story.[4] The point is that universality *amongst us*, individually and collectively, is quite enough to explain our sense that the direction of causation is an a priori matter – it is a priori, if you are built as we are, and if talk of causation depends on agency, in the sense that Ramsey suggests. (If Ramsey is right, then, the relevant contingency is not, as Russell thought, that memory only works backwards, but rather that deliberation only works forwards.)

[4]See Price (1996) for an introduction to the issue.

3.4 The Past Is Off-Limits . . . or Is It?

In making agency central in this way, in spotting the special character of the agent's epistemic perspective, and in seeing the work it can do in giving us an account of the difference between cause and effect, I think Ramsey gets much further than Russell. But there's one respect in which Russell still outdoes him, in my view. Ramsey actually gives us no reason at all why the epistemic bifurcation he identifies – the split between things for which our present volition can count as evidence, and things for which it can't – should line up neatly with the future/past distinction. He asserts confidently that it does, or at least that the entire past lies on the latter side of the line: "What is true is this, that any possible present volition of ours is (for us) irrelevant to *any* past event." But he offers no argument at this point, or explanation of the fact in question, if it is a fact. Russell at least has a punt at it, associating it with "the fact that the past but not the future can be known by memory."

This part of the puzzle wasn't properly sorted out until the 1960s, when the Oxford philosopher Michael Dummett pointed out (Dummett 1964) that one *could* quite coherently take oneself to affect the past – that is, in Ramsey's terms, take one's present volitions to be relevant to some past event – so long as one didn't take oneself to be able to *know* about the relevant part of the past, before one acted.[5]

Dummett points out that the knowledge condition might coherently be held to fail: there might be past events of which we thought that we couldn't have knowledge, at least at present, even in principle. If so, it isn't incoherent to think that we might affect them, by bring about some state of affairs in the future, with which they in turn are reliably correlated. Our initial action has to lie in the future, of course – that's the point above, about the turtles needing to swim East to get away from the beach. But in Dummett's picture our influence can then zig-zag back into the past, much as our turtles might then swim West, if gaps or inlets in the beach permitted it. In this respect then, I think that Ramsey misses something important. He's right about the character of the difference between cause and effect, but wrong about how it needs to line up with the distinction between past and future. Dummett shows how it might be false that, as Ramsey puts it, "any possible present volition of ours is (for us) irrelevant to any past event."

Dummett's own interest was in the efficacy of retrospective prayer, but I think his ideas have a more down-to-earth application, in the case of quantum mechanics – the reason we haven't noticed the gaps in the beach being that they so small, as it were. This is admittedly not a popular view, and I don't have time to explain it here, but I would like to put on record the view that the reason it has remained unpopular is in part because people who think about such things are unaware of the deep insight of Ramsey's view of causation, that it is the agent who is in the driver's seat, in determining the direction of causation in nature (in so far as there is such a

[5]Interestingly, Dummett's discussion can be taken to show how the usual objections to backward causation rely on the same tension between knowledge and free action that lies at the heart of Ramsey's proposal – see Ahmed and Price (2012), Sect. 3, Price (2012).

thing), not the other way round. Or, if they are aware of it, they haven't also noticed what Ramsey got wrong and Dummett got right, that the resulting causal arrow need not point exclusively to the future. It may also be that Dummett's loophole is particularly hard to see, because our ordinary ways of causal thinking are so deeply ingrained – because they are hardwired, in effect. There has been much fascinating work in recent years by psychologists such as Alison Gopnik (e.g., Gopnik 2009), revealing the extent to which our causal thinking is innate (developing in various stages during infancy). If one were designing a folk physics for creatures like us, there's no doubt at all that simply treating the past as fixed – saying "Don't even think about affecting the past!" – would be a useful place to start. It's quick, and even if I'm right it's only dirty in places that are never going to matter, in ordinary life. (Similarly, turtles might be hardwired to swim East, even in an environment in which there were occasional opportunities to swim West.)

3.5 Summary: What We Get from Ramsey

Setting aside this last issue, we can see that Ramsey offers us the beginnings of an answer to the two big puzzles about causation we identified earlier:

1. *Field's challenge*. Ramsey shows us how to reconcile Cartwright with Russell, without hyper-realism. Ramsey's Special Agent gives us the distinction Cartwright shows us that we need, between probabilistic dependencies encoded in the objective laws of association, and those that survive from the point of view of the deliberating agent.
2. *The direction of causation*. The difference between cause and effect is accounted for in terms of what we can manipulate to do what, and the prevailing temporal orientation of this "causal arrow" is explained in terms of a contingent asymmetry in us, the fact that we deliberate "past-to-future".[6]

And the upshot is that Russell comes off rather well – if Ramsey is right, so too is Russell, about the claim that causation is not as fundamental as many philosophers have thought. More on this in a moment, but first a word about counterfactuals.

4 A Home for Counterfactuals

Where do counterfactuals fit in? A couple of pages earlier in "General Propositions and Causality", Ramsey has this to say about the kind of conditional judgements we need in practical deliberation:

[6]Itself traceable presumably to the thermodynamic asymmetry, though I haven't said anything about that here. See Price and Weslake (2009) and Price (2007).

> When we deliberate about a possible action, we ask ourselves what will happen if we do this or that. If we give a definite answer of the form 'If I do *p*, *q* will result', this can properly be regarded as a material implication or disjunction 'Either not-*p* or *q*.' But it differs, of course, from any ordinary disjunction in that one of its members is not something of which we are trying to *discover* the truth, but something it is within our power to *make* true or false. (Ramsey 1929, p. 142, emphasis added)

Ramsey adds a footnote at this point: "It is possible to take one's future voluntary action as an intellectual problem: 'Shall I be able to keep it up?' But only by dissociating one's future self." (*ibid.*) What he is emphasising here is the special epistemic character of the agent's perspective, whose role a couple of pages later in the paper I described in Sect. 3.2 above.

Ramsey then continues:

> Besides definite answers 'If *p*, *q* will result', we often get ones 'If *p*, *q* might result' or '*q* would probably result'. Here the degree of probability is clearly not a degree of belief in 'Not-*p* or *q*', but a degree of belief in *q* given *p*, which it is evidently possible to have without a definite degree of belief in *p*, *p not being an intellectual problem.* And our conduct is largely determined by these degrees of hypothetical belief. (*ibid.,* emphasis added)

In other words, Ramsey is claiming here that the kind of conditionals we need in deliberation are not counterfactuals, not claims about what *would* be the case had something been different. They are hypotheticals, claims about what *is* true if something is the case, for a very special kind of "something" – propositions that can't be intellectual problems, as he puts it, because we take them to be "within our power to *make* true or false."

If Ramsey is right, then we don't need counterfactuals, where many people have thought that we do need them, as a foundation for our theory of decision. Instead we have the prospect that we might be able to explain talk of counterfactuals in terms of these simpler kind of hypotheticals. Something like this idea has been proposed by the psychologist Alison Gopnik, I think (see, e.g., Gopnik 2009). Her idea is that we can employ these deliberative skills not merely online, when facing real choices, but also offline, when facing imaginary choices (or, perhaps a better way to put it, when *imagining* that we are facing real choices). "Counterfactuals are the price we pay for hypotheticals", as Gopnik puts it: they are a kind of by-product of the hypotheticals we need in decision, combined with our faculty of imagination. Again, I don't have time to explore this idea here. The point I want to make is simply that if it is correct, then it will give us a satisfyingly Russellian account of counterfactuals, just as Ramsey's main proposal does for causation.

5 Causal Republicanism

So are there causes, if Ramsey is right? I think the answer is, "Yes, but they're not as much part of the furniture as we might have thought." Causation goes on the side of the secondary qualities, to use an analogy I explored in an early paper with Peter Menzies (Menzies and Price 1993). As Richard Corry and I noted in a

recent collection (Price and Corry 2007), we can characterise this option in terms of Russell's own constitutional metaphor. In the political case, we can distinguish three views of political authority. A traditional monarchist, at one extreme, takes it to be vested in our rulers by God. If we reject that view, we have two choices: there's the *anarchist* option of rejecting the notion of political authority altogether; or the milder *republican* option, which agrees with the traditional monarchist that there is political authority, but sees it as a social creation, vested in our rulers by us.

By analogy, the republican option exists in metaphysics, too. Causal republicanism is thus the view that although the notion of causation is useful, perhaps indispensable, in our dealings with the world, it is a category provided neither by God nor by physics, but rather constructed by us. From this republican standpoint, then, thinking of eliminativism about causality as the sole alternative to full-blown realism is like thinking of anarchy as the sole alternative to the divine right of kings.

As I noted, there's an issue about where we put Russell: is he an anarchist or a republican, in this new terminology? I suggested that we might see him, with Hume, as what now counts as an armchair anarchist, happy to retain talk of causation and counterfactuals for ordinary purposes. And that's really the republican option, or at least the reflective, clear-headed version of the republican option, that wants to combine a full-blooded participation in ordinary ways of speaking with a detached, over-the-shoulder understanding of how we come to speak that way, and how it might have been otherwise – "contingency, irony and solidarity," in the words of another of my philosophical heroes.

How is this news going to be taken by the market? It won't be a shock, presumably, to learn that market values reflect the subjective preferences we humans have for various possible ends and outcomes. (Another of Ramsey's great contributions was to figure out how to systematise and measure such things, along with our degrees of belief, in the context of their role in guiding our choices.) All of this depends on the fact that we are creatures who operate under uncertainty: "All our lives we are in a sense betting", as Ramsey puts it (1926, p. 85). What the republican view of causation does is to extend the story to the case of creatures who are also *agents* – they *intervene* in the environment about which they hold such beliefs and preferences. Certainly it shows that causal and counterfactual talk would be unneeded for creatures who didn't do that – intelligent trees, for example, as Michael Dummett (1964, p. 339) once put it. But that's no threat to the real estate market, obviously, unless the good folk of Manhattan become a lot more passive than they have tended to be to date.

6 Russell and the Monarchy

So much for causation. But when it occurred to me that I had an opportunity to celebrate the centenary of "On the Notion of Cause" on this occasion, I was curious about what Russell had had in mind in the other part of his famous line: just *what*,

in Russell's view, is the harm that the monarchy is erroneously thought not to do? I assumed that this would be an easy curiosity to satisfy – somewhere, presumably, Russell would have expressed his views about the monarchy at greater length. But I searched in vain.

Eventually I wrote to Nicholas Griffin, of the Russell Archives at McMaster. He told me that there was really nothing to find, not even in Russell's correspondence, so far as he knew it. But he did suggest a context for Russell's remark. In 1910 Britain had concluded a considerable constitutional crisis, bought on by the Liberal government's determination to remove the veto power of the House of Lords. A crucial step had been the King's indication that he would support the government, if necessary, by creating a sufficient number of new Liberal peers to ensure passage of the Bill through the Lords. (Russell himself would have been one of those new peers, apparently, in that counterfactual world.) Professor Griffin suggested that in the light of the King's support, some on the Liberal side of politics were inclined to say that the monarchy wasn't so bad after all; and that Russell may have been taking the opportunity to indicate that he was made of sterner stuff – that the old battle lines of the Russells remained unchanged, as it were.

But that doesn't tell us what Russell thought that the harm in question actually was, at that point in the nation's history – when, thanks in part to Russell's own ancestors, it had long been a "crowned republic", as Tennyson put it (a fact reaffirmed and strengthened in the recent crisis, of course). So, as my centenary footnote to Russell's great paper, I want to finish by giving you my own view. In my view, there is a significant cost to modern constitutional monarchies that is remarkable, among other things, for the fact that although it is in plain sight, it goes unmentioned, and apparently almost unnoticed. As you'll see, it is indeed a relic of a bygone age, whose significance is hidden from us by the sheer familiarity of the system of which it is a consequence – by the fact that a traditional picture holds us in its grip, as Wittgenstein might have put it. Moreover, while I'm not suggesting that this is what Russell actually had in mind, it is, as you'll also see, something that he in particular would have had reason to have in mind – it resonates in several ways with significant aspects of his own life. And it connects in a deep way with the themes I have been talking about so far. In all senses, then, it's an excellent fit.

I can introduce the point by noting a difference of opinion with most of my Australian compatriots. A majority of Australians appear to favour an Australian republic, even though many of them voted against it when they had the chance some years ago, because they didn't like the model on offer. But the main reason given is that Australia should have an Australian head of state, rather than the British monarch – and while I don't disagree with that sentiment, I do think it misses a much more powerful appeal to the country's professed values.

Australia's next head of state, under present arrangements, is a man whose affection for the country dates from 1966, when he was sent for a couple of terms to a boarding school there. It was a formative experience, apparently – "If you want to develop character, go to Australia" (Wales 2011), as he put it recently (going on to mention some of the character-building epithets employed by his Australian schoolmates). I feel very much on the same wavelength as the Prince at this point,

because I encountered those same character-building opportunities, and no doubt the same epithets, the very same year, as a teenage migrant from the UK – I arrived just three weeks after he left.

Our lives diverged quite markedly after that point, of course. He came to Cambridge, had a memorable gig in Wales in 1969, and has now served his country and the Commonwealth with considerable distinction, for more than 40 years. I went to ANU, making the first of many choices that turned out to lead, happily and rather surprisingly, to the present occasion. (Not quite a gig in Wales, perhaps, but memorable for me, nonetheless!)

But the particular difference I want to highlight is that in common with most of my generation, in countries such as Britain and Australia, I made choices about what to do with my life; whereas he did not, to an unusual extent. Significant as his life's work is, my famous contemporary did not have the opportunity to *choose* it, or to volunteer for it, in any meaningful sense.

So that's why I disagree with most of my compatriots, republicans and monarchists alike, about the issue of an Australian republic. They think that the question turns on the importance or otherwise of Australia's having an Australian head of state. I think that that's a trivial matter, a mere sideshow, compared to the principle that all young people should be allowed to choose for themselves what they do with their lives, when they grow up. The professed Australian value I mentioned a moment ago is simply that of fairness – of a "fair go", as Australians say. It seems to me profoundly and manifestly unfair to select children by accident of birth for future public office – especially so, of course, for such an important, symbolic and *public* public office – and hence entirely inappropriate that Australia's constitution should make us party to a practice of doing so.

Apart from the fact that Australians like to make a fuss about fairness, there's nothing uniquely Australian about the point, of course. It applies with equal force in all the modern democratic monarchies – almost all of which (eight in total, neglecting some tiny principalities) are in Europe. The eight countries in question – three Scandinavian countries, three Benelux countries, Spain, and Britain – actually care just as much about fairness as Australians do, of course. So it's an issue for all of them, in just the same way.

It is easy to see why I think this point is Russellian in spirit. As is well known, Russell felt the constraints of his own childhood very deeply, and was greatly relieved to escape them when he came of age.[7] Later, when he himself became a father, Russell was a famous advocate of allowing children as much freedom as possible. And finally, of course, he was also a famous opponent of conscription.

It would be a little extreme, perhaps, to compare hereditary monarchy to conscription. A neutral term might be "involuntary service", or "involuntary servitude", as prohibited by the landmark Thirteenth Amendment to the US Constitution.

[7]The heirs to the monarchies of Europe don't have that opportunity, of course. It is true that in principle they could abdicate, but at considerable cost – like it or not, they are public figures, after all – and only by passing the unasked-for obligation to a sibling or cousin.

In that case, the US Supreme Court later ruled that the Amendment did not prohibit service properly rendered to the state, such as military conscription, or jury duty. But as those cases make clear, it would be unthinkable that these exclusions might be lifetime matters (or apply to individuals chosen in infancy, presumably).

It would be unthinkable in any other advanced democracy, too, if we were starting from scratch. If someone proposed that we should fill public offices by selecting infants who would be brought up to fill the roles in question, the main objection would not be that it was undemocratic, but that it was absurdly unfair to the individuals concerned. The fact that we do find this system thinkable in practice turns mainly on its sheer familiarity – that's just how things are done. And perhaps, as Russell thinks in the case of causation, we are still in the grip of a piece of bad metaphysics: we think of royalty as a natural kind, and hence imagine that it is a natural matter that royal children should be brought up to play these roles – that's the kind of beings they are, as it were. The picture holds us captive, and central to it is the fantasy that what these families enjoy is a matter of entitlement and privilege, not constraint and obligation.

It is easy to see how we got to this point, from the distant past this picture actually depicts: on the one hand, a great erosion of power on the side of royalty, as – thanks in part to Russell's ancestors, in the British case – its powers were curtailed; on the other hand, an even greater expansion of opportunity on the side of ordinary people, especially ordinary children, as we have come to accept that young people should make their life choices for themselves, rather than have them dictated by parents or accidents of birth. The combination of these two factors means that the heirs to modern monarchies are now marooned on a little island of underprivilege: impoverished not only compared to their own ancestors, but also, much more importantly, by the standards that now exist in the community at large.

6.1 Counterfactual Deprivation

Here the point connects back to the main themes of my lecture, and indeed of Russell's, broadly interpreted. For what precisely is it that these individuals lack, compared to their contemporaries? Of what are they deprived? Essentially, it is counterfactuals. They are deprived of counterfactuals, of opportunities to make decisions about their lives on the large scale, in much the same way that New Jersey is deprived compared to Manhattan, about admittedly more trivial matters, to return to Fodor's observation about the real estate market.

So the two sides of Russell's famous metaphor come together. The reason we need to be at most armchair anarchists about causation and counterfactuals, turns, in part, on the role these notions play in human life, and particularly our conception of a rich human life, a life with choices, a life with counterfactuals. It would be unthinkable, probably impossible, to live life without them, even if they are not fundamental, and have no role in fundamental physics, suitably construed. And the great flaw of modern hereditary monarchies lies in the way in which they deprive

a few individuals of some of these freedoms. This flaw has precisely the character suggested by Russell's comparison: a significant harm, in plain sight, to which we are blinded by familiarity and bad metaphysics.

7 Conclusion

With some renovation, then – new foundations provided by Ramsey on the side of causation, and a sympathetic addition of the wing that Russell himself did not construct, on the side of the monarchy – Russell's edifice thus turns out to be in remarkably good shape, on the eve of its 100th birthday. Please join me in wishing it well.

References

Ahmed, A., and H. Price. 2012. Arntzenius on 'Why Ain'cha rich?'. *Erkenntnis* 77: 15–30.

Anscombe, G.E.M. 1971. *Causality and determination*. Cambridge: Cambridge University Press. Reprinted in her *Metaphysics and the philosophy of mind: Collected philosophical papers volume III*, 133–147. Minneapolis: University of Minnesota Press, 1981. Page references are to the latter version.

Cartwright, N. 1979. Causal laws and effective strategies. *Noûs* 13: 419–437.

Collingwood, G. 1940. *An essay in metaphysics*. Oxford: Oxford University Press.

Dummett, M.A.E. 1964. Bringing about the past. *Philosophical Review* 73: 338–359.

Field, H. 2003. Causation in a physical world. In *Oxford handbook of metaphysics*, ed. M. Loux and D. Zimmerman, 435–460. Oxford: Oxford University Press.

Gasking, D. 1955. Causation and recipes. *Mind* 64: 479–487.

Gopnik, A. 2009. *The philosophical baby: What children's minds tell us about truth, love, and the meaning of life*. New York: Farrar, Straus and Giroux.

Lewis, D. 1979. Counterfactual dependence and time's arrow. *Noûs* 13: 455–476.

Mellor, D.H. 1988. *The warrant of induction*. Cambridge: Cambridge University Press. Reprinted in his *Matters of metaphysics*, 254–268. Cambridge: Cambridge University Press, 1991. Page references here are to the latter version.

Menzies, P., and H. Price. 1993. Causation as a secondary quality. *The British Journal for the Philosophy of Science* 44: 187–203.

Pearl, J. 2000. *Causality*. New York: Cambridge University Press.

Price, H. 1991. Agency and probabilistic causality. *The British Journal for the Philosophy of Science* 42: 157–176.

Price, H. 1993. The direction of causation: Ramsey's ultimate contingency. In *PSA 1992, Volume 2*, ed. D. Hull, M. Forbes, and K. Okruhlik, 253–267. East Lansing: Philosophy of Science Association.

Price, H. 1996. *Time's arrow and Archimedes' point: New directions for the physics of time*. New York: Oxford University Press.

Price, H. 2007. Causal perspectivalism. In *Causation, physics, and the constitution of reality: Russell's republic revisited*, ed. H. Price and R. Corry, 250–292. Oxford: Oxford University Press.

Price, H. 2012. Causation, chance and the rational significance of supernatural evidence. *Philosophical Review* 121: 483–538.

Price, H., and R. Corry (eds.). 2007. *Causation, physics, and the constitution of reality: Russell's republic revisited*. Oxford: Oxford University Press.

Price, H., and B. Weslake. 2009. The time-asymmetry of causation. In *The Oxford handbook of causation*, ed. H. Beebee, C. Hitchcock, and P. Menzies, 414–443. Oxford: Oxford University Press.

Rabinowicz, W. 2002. Does practical deliberation crowd out self-prediction? *Erkenntnis* 57: 91–122.

Ramsey, F.P. 1926. Truth and probability. In *Foundations: Essays in philosophy, logic, mathematics and economics*, ed. D.H. Mellor, 58–100. London: Routledge and Kegan Paul.

Ramsey, F.P. 1929. General propositions and causality. In *Foundations: Essays in philosophy, logic, mathematics and economics*, ed. D.H. Mellor, 133–151. London: Routledge and Kegan Paul.

Russell, B. 1913. On the notion of cause. *Proceedings of the Aristotelian Society, New Series* 13: 1–26.

Suppes, P. 1970. *A probabilistic theory of causality*. Amsterdam: North-Holland.

von Wright, G. 1973. On the logic and epistemology of the causal relation. In *Logic, methodology and philosophy of science IV*, ed. P. Suppes, L. Henkin, G.C. Moisil, and A. Joja, 293–312. Amsterdam: North-Holland.

von Wright, G. 1975. *Causality and determinism*. New York: Columbia University Press.

Wales, C. 2011. Personal remarks cited in: 'Pommy Bastard' Prince Charles reveals 'Huge Affection' for Australia. *The Age*, 27 Jan 2011. Online at http://www.theage.com.au/world/pommy-bastard-prince-charles-reveals-huge-affection-for-australia-20110127-1a5na.html. Accessed 8 Apr 2013.

Woodward, J. 2003. *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Pragmatism and European Philosophy: William James and the French-Italian Connection

Massimo Ferrari

1 James and Renouvier

"J'ai commencé à lire votre *Pragmatism* au moment où la poste me l'a remis et je n'ai pas pu le déposer avant d'en avoir achevé la lecture. C'est le programme, admirablement tracé, de la philosophie de l'avenir". In this letter of June 27, 1907 Henri Bergson expressed his great admiration for the American philosopher William James.[1] Two years before Bergson had already noted the surprising affinities between American pragmatism and new French philosophy: two philosophical perspectives, in his mind, which were flourishing in totally independence of each other, but ultimately destined for full agreement.[2] In the early years of twentieth century James was one of the more appreciated and well-known thinkers within French philosophy, whereas other eminent figures of the pragmatist movement were not influential. Charles Sanders Peirce's writings were not easily available, although his seminal essay *How to Make Clear our Ideas* was already translated into French and published in 1879 in the *Revue Philosophique*.[3] John Dewey seemed to be still a scholar at the beginning of his career, whereas Ferdinand Schiller was more discussed and known also thanks to his participation at the 3rd International Congress of Philosophy in Heidelberg (1908). But the pragmatist *par excellence* in France was doubtless William James, though his reception was sometimes influenced by the commonplace of a Yankee philosopher embodying the business mentality typical of American civilization. Nevertheless – and this was very

[1] (Bergson 2011, p. 74).

[2] *Id.*, p. 73 (letter to James of July 20, 1905).

[3] (Peirce 1879).

M. Ferrari (✉)
Department of Philosophy, University of Turin, Via S. Ottavio 20, 10124 Turin, Italy
e-mail: massimo.ferrari@unito.it

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_43,

different from what happened in neighbouring Germany – James quickly became a very respected philosophical figure and it was precisely Bergson who established James as the unavoidable partner for the new philosophy opening the twentieth century.

The relationship between James and French philosophy nevertheless was rooted in a longer past. James' apprenticeship was deeply influenced by Charles Renouvier, the founder of a French Neo-Kantian tradition. Since his early philosophical studies James was well acquainted with his thought and he contributed later to the journal edited by Renouvier and François Pillon which first appeared in 1867 with the title *L'Année Philosophique*, but since 1872 until 1889 was well known in the philosophical community as *La critique philosophique*.[4] The correspondence and the intellectual exchange between James and Renouvier was later accurately analysed by Jean Wahl who, in a seminal essay first published in 1922 about the two volumes containing the correspondence of James, gave a vivid picture of James's encounter with Renouvier's philosophical works as well as of the famous occasion in which James, by reading Renouvier, discovered for the first time the cornerstone of his pragmatist view. According to James, it was a "crucial day" in his life when he understood thanks to Renouvier's second volume of his *Essay de critique générale* that "the first act of free will consists exactly in believing in free will".[5] The story told by Wahl in his influential book, and later described by Ralph Barton Perry in his wonderful reconstruction of James's intellectual biography, deals not only with James's early acquaintance with Renouvier and the philosophical climate of late nineteenth century in France, but also casts a new light about the "continental" roots of pragmatism.[6]

To be sure, as Perry stated, Renouvier had "the greatest single influence on James's thought".[7] Pluralism, indeterminism, the postulates of practical reason and – to some extent – a kind of Kantianism are the essential topics James inherited from Renouvier, though in his mind the philosophy professed by the French thinker can be interpreted within a more general frame of empiricism.[8] According to the young James, Renouvier's main point was the conception of freedom, which he

[4]James was a reader of the *Année Philosophique* already in 1867. In its first volume it offered Pillon's programmatic editorial, where he made a plea for the principles of Kantian philosophy, but in the sense of a Kantianism free from the dialectical antinomies (Pillon 1867). Renouvier's extensive essay in the same volume was undoubtedly very stimulating for James, in particular for his conclusion stressing the "revolution" accomplished by Kant also in the field of moral philosophy ("une révolution dans les esprits et dans les coeurs, et une revolution morale nécessaire"). See (Renouvier 1867). What also must be remembered is that some of James's essays (among them the famous *Sentiment of rationality*) were translated into French by Renouvier himself.

[5](Wahl 2004, p. 58).

[6](Perry 1936). Perry had already edited – between 1929 and 1932 – the letters of William James and Charles Renouvier (Perry 1929, 1935).

[7](Perry 1936, vol. I, p. 465).

[8](Wahl 2004, p. 71).

had elaborated "for the first time" in a very "intelligible and reasonable" fashion.[9] This was a crucial aspect of Renouvier's Neo-Criticism and James could find a clear formulation of such a point of view in that second volume of Renouvier's *Essais de critique générale* he had read at the very beginning of his philosophical development. Renouvier was convinced that every act of the will has to face the question "to be or not to be?", and that this definition of the will at the same time defines what is freedom. The will – Renouvier argued – is in no way related, quite differently from spontaneity, to the necessity of the world. Following this main idea Renouvier refuted the Kantian antinomies and the division of men's reason in two parts, vindicating by contrast the essential *unity* of theoretical and practical reason. This unity was the unavoidable premise, Renouvier suggested, for our *belief* in human freedom, and this was possible only assuming a strict phenomenalism excluding once for all any kind of thing in itself.[10]

By reading these pages of Renouvier's outstanding work James incurred a profound debt towards his French colleague. Thanks to Renouvier, pluralism, determinism, and freedom, but also a very peculiar sort of Kantianism became essential elements of James's philosophy.[11] Surely James and Renouvier didn't agree about some pivotal points and, in particular, Renouvier was very sceptical towards James's idea of a stream of consciousness. For Renouvier it seemed quite impossible to create a science of psychology "without recognizing an intellectual basis for such general terms as *where*, *who*, *when*, *what*, *for what*, *by what*". This was precisely the peculiar heritage of his way to interpret Kant's theory of categories which he claimed to be still valid and, at the same time, why he found it very difficult to comprehend James's psychological inquiries. By contrast, the new type of psychological orientation elaborated in his *Principles of Psychology* was for James exactly the perspective we have to accept in place of our ordinary psychological presumptions.[12] That disagreement notwithstanding, Renouvier remained for James one of his most important philosophical partners. He said, in a letter to Renouvier on August 4, 1896 that his article about the *Will to believe* was a sequel to Renouvier's philosophy: by reading it, James added, "you probably recognized how completely I am still your disciple".[13] Indeed this "disciple" dedicated later to Renouvier his uncompleted book *Some Problems of Philosophy*, recalling the "decisive impression made on [him] in the seventies by his masterly advocacy of pluralism", which had freed him "from the monistic superstition under which [he] had grown up".

[9] See the letter to Renouvier written on November 2, 1872 (Perry 1936, vol. I, pp. 661–662).

[10] (Renouvier 1859, pp. 207, 409–413).

[11] About James's Kantianism see the illuminating contribution (Carlson 1997, pp. 363–383; on James and Renouvier see especially p. 365).

[12] We refer to an exchange of letters between James and Renouvier (from September 1884 to March 1887) published in (Perry 1936, vol. I, pp. 668–702).

[13] *Id.*, p. 709.

Still shortly before his death in 1910, James regarded Renouvier as "the strongest philosopher of France during the second half of the twentieth century".[14]

2 Pragmatism and French Philosophy

James's intellectual commitment to Renouvier represents a basic aspect of his relationship with French philosophy and his reception in France until the decisive encounter with another great exponent of it, namely Henri Bergson. First and foremost we must emphasize that the early reception of pragmatism in French culture before World War I was characterized by the discussion of James as the eminent spokesman of a new philosophy and, primarily, as the author of the *Varieties of Religious Experience* (promptly translated in French in 1906). This aspect is particularly worthy of attention. James was well accepted in the circles of religious modernism, which were flourishing in France at that time. Modernism was not only devoted to renewing Catholicism through the contact with modern culture, historical knowledge, and a vehement polemic against the revival of Thomism, but it also agreed that it was the practical consequences of ideas that served as a criterion of their truth. One of the fountainheads of French Modernism was Maurice Blondel, who first labelled his "philosophy of action" devoted to give full freedom to spirit and spontaneity of faith as "*pragmatisme*". For a short but highly important time, pragmatism and modernism were close allies therefore – a circumstance that we can find also in Italy, in particular by considering the monthly journal *Il Rinnovamento*, organ of Italian modernism between 1907 and 1909.[15]

The core of James' increasing fortune in France lay in his encounter with its academic philosophy, namely Henri Bergson on the one hand and Émile Boutroux on the other. Both were personally and philosophically very well acquainted with James and James, for his part, had no hesitation in recognizing the great merits of his French colleagues. Since 1902 he was an admirer of Bergsons's *Matière et mémoire*, though he was aware of some notable differences concerning Bergson, in particular, the function of the unconscious in our mental life.[16] Their agreement about the two cardinal points of their psychological inquiries (the stream of consciousness and the *durée réelle*) was, on the other hand, not as evident as has long been suggested, given the fact – clearly underlined by Bergson – that the *Principles of Psychology* and the *Essai sur les données immédiates de la conscience* arose as two

[14](James 1979, pp. 84–85).

[15]For an exhaustive account of the relationship between pragmatism and modernism in France at the beginnings of twentieth century see (Shook 2009). Regarding similar aspects in Italian culture and the review *Il Rinnovamento* we refer to (Ranchetti 1963, pp. 191–226) and (Scoppola 1961, pp. 163–220).

[16]We refer to the letters available in (Perry 1936, vol. II, pp. 606, 609) and in (Bergson 2011, p. 63).

independent and original works.[17] Moreover, other supposed philosophical affinities have to be considered in the light of the different ways in which James and Bergson elaborated their radical empiricism and conceived of the relationship of intelligence to reality.[18] Despite these and other problematic features characterizing James's and Bergson's understanding of their respective philosophies, it seems undeniable that James saw in Bergson the most ardent supporter of his fight against intellectualism, and Bergson, for his part, considered James, since his first acquaintance with him (namely with the *Varieties of Religious Experience*), a very profound and impressive thinker.[19] Furthermore, James's *Pragmatism* and Bergson's *Évolution créatrice* appeared even by a symbolic coincidence in the same year: for both the American pragmatist and the new hero of French philosophy 1907 was a veritable *annus mirabilis*. James's enthusiastic reception of Bergsons's work speaks for itself: "Your new book is just arrived – hurrah! hurrah! hurrah!". And a month later: "O my Bergson you are a magician, and your book is a marvel, a real wonder in the history of philosophy . . . The vital achievement of the book is that it inflicts an irrecoverable death-wound upon Intellectualism".[20]

No wonder also, that James, in his lecture on Bergson ("Bergson and his Criticism of Intellectualism" published 1909 in the *Pluralistic Universe*), underlined that Bergson's was the first "radical" critique of every form of rationalism and intellectualism. According to James, Bergson could be compared with Kant and his successors. Whereas Kant's conception of knowledge was founded on a variety of data synthesized by reason through its immutable intellectual forms, Bergson was by contrast convinced that "the flux of life undergoes only at our hands in the interests of practice . . . ; and to understand life by concepts is to arrest its movement, cutting it up into bits as if with scissors, and immobilizing these in our logical herbarium".[21] According to James, this was the outcome of a new conception of reality, which is never completed, but only ever in the making: "What really *exists* is not things made but things in the making. Once made, they are dead, and an infinite number of alternative conceptual decompositions can be used in defining them". So James was even a little emphatic by stressing such a novelty of Bergson's philosophy. It was –

[17] In a letter to James of May 9, 1908 Bergson noted that he started from a great interest in physics and mathematics. So his discovery of *durée réelle* had nothing to do with the approach to conception of time James proposed a year later in his *Principles*. "Ce fut l'analyse de la notion de temps telle qu'elle intervient en mécanique ou en physique, qui bouleversa toutes mes idées" (Bergson 2011, p. 77).

[18] A very useful overview of Bergson's and James's philosophies, their similarities and differences is available in (Madelrieux 2011). See also (Worms 1999) for an illuminating comparison between the two philosophers.

[19] (Bergson 2011, p. 58): letter to James of January 6, 1903.

[20] We refer to the letters James wrote to Bergson on May 19, 1907 and June 13, 1907 published in (Perry 1936, vol. II, pp. 618–621). In a letter to Charles Strong of the same period James remarked: "Have you read Bergson's *Évolution créatrice*? It seems to me the absolutely *divinest* book on philosophy ever written up to this date" (*id.*, p. 604).

[21] (James 1996, p. 244).

he said – "like the breath of morning and the song of birds".[22] Put in more conceptual words, James greatly appreciated the new horizon opened by Bergson due to his capacity to overcome the traditional view of *logos* and discursive thought. It was indeed a "revolution" aiming at a concept of truth resting on raw unverbalized life.[23]

Dear James, said Bergson for his part from Paris, we totally agree.[24] According to Bergson, the conflict between intellectualist positivism and pragmatism was a decisive one. Quite differently from positivism and its faith in the given, pragmatism signifies "une conception nouvelle de la relation de l'abstrait au concret, et de l'universel à l'individuel. Pour ma part, je n'ai jamais mieux saisi l'affinité qui existe entre nos deux méthodes de penser. Nous avons en tous cas les mêmes adversaires et, comme vous vouliez bien me l'écrire il y a quelques mois, 'we are fighting the same fight'."[25] But Bergson also stressed some points of disagreement between them. A year before, in a letter of June 27, 1907, Bergson raised an interesting objection by asking whether the truth is mutable – as James suggested – or whether reality is something always changing (as Bergson seemed to argue).[26] Nevertheless, in his "Preface" to the French translation of James' *Pragmatism* published in 1911, Bergson endorsed James' pragmatism and gave his support to the pragmatist point of view. But in doing so, Bergson also stressed, on the one hand, a similarity between James and Kant. In his opinion, pragmatism had to be considered a sequel to Kantianism, due to the fact that truth depends for Kantianism on the universal structure of our mind, to which view pragmatism "adds, or at least implies, that the structure of the human mind is the effect of the free initiative of a certain number of individual minds".[27] On the other hand, Bergson regarded truth, in agreement with James, as an invention and not as a discovery of something readymade, but he emphasized at the same time, differently from James, that the truths of feeling and scientific truths are not of the same kind. Their difference is, so to speak, similar to the difference between a sailboat (*bateau à voiles*) and a steamer (*bateau à vapeur*). Both are human inventions, but the former makes only slight use of artificial means and takes the direction of the wind, whereas the latter needs an artificial and elaborate mechanism in order to sail in the direction which we ourselves have chosen.[28]

---

[22] *Id.*, p. 263.

[23] *Id.*, pp. 272–273.

[24] Bergson's great appreciation of James's essay devoted to him is well documented by the response he gave on May 18, 1910 to an article appeared in the *Journal of Philosophy*, where the reliability of James's presentation of the French philosopher was questioned. By contrast, Bergson clearly stated that James had perfectly understood the essential features of his own thought ("je tiens l'interprétation de W. James pour parfaitement exacte"). See (Bergson 2001, pp. 384–388).

[25] (Bergson 2011, p. 75): letter to James of January 27, 1908.

[26] *Id.*, p. 74.

[27] *Id*, p. 11.

[28] *Id.*, p. 12.

We don't know James's answer to the distinction Bergson proposed between two different ways of conceiving truth as invention. The great philosopher died in August 1910, at the top of his success in his country as well as in Europe, particularly in France. Surely it is no accident that in the following year another French philosopher such as Émile Boutroux was the first to publish a book about James's life and work.[29] James had invited Boutroux as visiting professor in Harvard at the beginning of 1910. About his lectures James wrote in the review *Nation* of March 31, 1910 a brief, but highly illuminating article entitled "A Great French Philosopher at Harvard". James saw in Boutroux the leading figure of the reaction against the abstract philosophy and he appreciated especially his famous book *La contingence des lois de la nature*. According to James, Boutroux had achieved in contemporary philosophy of science by developing, via a notion of contingency quite different from that of chance, a new interpretation of science and its conceptual categories. "At the present day ... concepts like mass, force, inertia, atom, energy, are themselves regarded rather as symbolic instruments".[30] A consequence of this point of view was for James that all the scientific entities are in reality to be understood in the same sense in which "a statue is previous in its rock".[31] Moreover James praised the pivotal concept of *contingency* as "the element of spontaneity which characterizes human life", opening the possibility for "*many* futures" and for causality without necessity.[32] James recognized also without hesitation the great merit of Boutroux's lectures in Harvard: on the one hand he had shown that scientific concepts too ought not to be sterilized and separated from the "purposes of living reason"; on the other hand Boutroux had made a high plausible appeal "to the fullness of concrete experience". In his account of the Harvard lectures James consequently pointed out the association of Boutroux's philosophy with the names of Peirce, Dewey, Schiller, Bergson and, obviously, of James himself: "It is the real empiricism, the real evolutionism, the real pluralism; and Boutroux (after Renouvier) was its earliest, as he is now its latest, prophet."[33]

Nevertheless, it should be noted that Boutroux seemed to be more interested in other features of James' pragmatism, as it is already documented both by the "sympathetic preface" Boutroux wrote to the French translation of James' *Varieties of Religious Experience* and by his later essay about James's philosophy of religion.[34] To be sure, Boutroux had a great respect for the acknowledgement of human personality that James conceived as the core of religious experience. But at the same time he had a different view of the essence of religion itself, which in

[29] See (Boutroux 1911). A very useful account of the relationship between James and Boutroux is available in (Perry 1936, vol. II, pp. 560–569).

[30] (James 1978, p. 169).

[31] *Id.*, p. 170.

[32] *Ibid.* See also Boutroux's letter to James on June 27, 1907 in (Perry vol. II, p. 766).

[33] (James 1978, p. 171). Regarding Boutroux's commitment to Renouvier and Renouvier's positive attitude toward the philosophy of contingence see (Parodi 1919, pp. 168–169).

[34] (Perry 1936, vol. II, p. 561).

his mind was not reducible only to a subjective, sentimental experience, being by contrast grounded also on intellectual, dogmatic, and institutional aspects.[35] James did not agree with this interpretation and refused the label of subjectivism, arguing that an object is "inseparable from the consciousness of it".[36]

In his very vivid portrait of James Boutroux offered an extensive account of the *Principles of Psychology*, whose primary result in his eyes was the method of introspection which allowed direct experience of ourselves, of the individual whole we are living in our psychological *Erlebnis*.[37] But Boutroux praised the fundamental work of his American colleague also for another, still more important reason. James's psychological enquiries had opened the path to a veritable psychology of religion, which Boutroux considered the great innovation of pragmatism. *The Varieties of Religious Experience*, he stated, inaugurated a new direction of experience as well as an approach to a basic aspects of human existence.[38] In his opinion the objectivity of religious experience was the framework for every field of objectivity, including psychological or scientific reality.[39] For Boutroux, it was in the essentially "open" character of his philosophy that James's most fundamental contribution to contemporary philosophical debate consisted, especially concerning the connection between reason and will, between knowledge and action.[40] Nevertheless, by remarking that the objectivity of science and psychology does not represent the only kind of veridical experience, but rests also on religious and moral life, Boutroux aimed to show, like Bergson, that the pragmatist concept of truth was to some extent questionable. On the one hand, scientific truth was different from religious or moral truth; and on the other hand, Boutroux said, Bergson was right in stressing that intellectual knowledge is not the only way to a deeper understanding of ourselves, that is of our inner experience which can be grasped only through intuition.[41]

This was the image of James' pragmatism prevailing within French philosophical landscape. Stimulated by James, Bergson and Boutroux had written the agenda of contemporary philosophy, or more precisely – as Eduard Le Roy said – of a revolution in philosophy comparable to the Bergson's own.[42] In this context, another circumstance often ignored can be considered as a meaningful document of the French discussion about pragmatism. On Saturday, Mai 8 1908 outstanding members of the *Société Française de Philosophie* discussed the main aspects

[35] See (Boutroux 1908, p. 337).

[36] More precisely, in his letter to Boutroux (July 20, 1908) James emphasized: "I am not, *epistemologically*, a subjectivist, in spite of what I call my radical empiricism; and I hold with you that faith in the existence of an object is inseparable from the consciousness of it, so far as that consciousness is effective in our life" (Perry 1936, vol. II, p. 563).

[37] (Boutroux 1911, pp. 30–31).

[38] *Id.*, p. 12, 52.

[39] *Id.*, p. 93.

[40] *Id.*, p. 142.

[41] *Id.*, pp. 89–90, 93.

[42] (Le Roy 1913, p. 13). On Le Roy see (Hill 2009).

characterising the new philosophy: Léon Brunschvicg, George Sorel, Dominique Parodi, Jules Tannery, René Berthelot, Eduard Le Roy and many others attended at this meeting, shortly before the more famous Congress of Philosophy in Heidelberg in September 1908 at which the *querelle* concerning pragmatism broke out in Germany.[43] The point of view endorsed especially by Le Roy in Paris consisted in recognizing a version of pragmatism he labelled as "French pragmatism". Dominique Parodi was convinced that pragmatism had already arisen in France thanks to the contributions of Lucien Laberthonnière and Maurice Blondel, whose philosophy of action showed an affinity – as already suggested above – with the new American way of thinking.[44] Nevertheless, Laberthonnière stressed in this context that pragmatism and the philosophy of action were quite different, given the fact that the former was engaged against metaphysics and the latter was, by contrast, grounded on the will "to elaborate a metaphysics" (*faire de la métaphysique*).[45] But it was in particular Le Roy who had a prominent role during the discussion. He attempted – only partly successfully – to suggest that James's pragmatism could be interpreted as a theory of knowledge and as a philosophy of experience. According to Le Roy this view was compatible with modern science and its development, once we have admitted that both our ordinary human intelligence and scientific intelligence are always changing.[46] In Le Roy's mind the crucial point seemed to be that scientific theories constitute first and foremost "tools useful for coordination and inquiry, characterized in part by an unavoidable choice, a conventional symbolism". In this sense pragmatism was right in claiming the usefulness of theories and of truth itself, but their point lay nonetheless in their scientific content and not merely in being utilitarian instruments at the heart of scientific explanation.[47]

So Le Roy was mainly interested in stressing an aspect widely neglected by his contemporaries, namely James's connection to philosophy of science.[48] Le Roy was not alone in this view. Few years later, in his book *Un romantisme utilitaire*, René Berthelot argued that pragmatism was not only similar in its origins to Nietzsche's radical refusal of traditional philosophy, but could be considered as a good partner

[43](Berthelot et al. 1908, pp. 249–296).

[44]*Id.*, p. 265.

[45]*Id.*, pp. 281–282.

[46]*Id.*, p. 276.

[47]*Id.*, p. 278.

[48]Interesting enough, a great scholar of Leibniz's logical work and fervent supporter of Bertrand Russell's logicism such as Louis Couturat strongly rejected pragmatism as a kind of Protagorean philosophy. He maintained that "le pragmatisme est le dernier avatar de l'empirisme; il en pousse à bout, non sans une certaine rigueur logique, les conséquences agnostiques et sceptiques". Without referring explicitly to James, Couturat resolutely opposed the "philosophy of action" and added in this sense: "À cette philosophie serve, quelle que soit sa forme, moralisme ou pragmatisme, nous opposerons la philosophie libre dont Descartes a formulé les règles et posé les fondements". See (Couturat 1983, pp. 28–31).

for scientists such as Poincaré.[49] Particularly, Berthelot suggested that Poincaré was a pragmatist (or a kind of "half pragmatist") in how he conceived of the validity of geometrical axioms, not as true *per se*, but exclusively with regard to their convenience.[50] Despite their undeniable interest, Berthelot's arguments were not enough in order to consolidate a different image of James' pragmatism in French philosophy.[51] In his *Pragmatism* James had indeed offered a short, but very illuminating account of contemporary philosophy of science. Mach, Duhem and Poincaré – James argued – were "teachers", according to whom "no hypothesis is truer than any other in the sense of being a more literal copy of reality. They are all but ways of talking on our part, to be compared solely from the point of view of their *use*".[52] Moreover, James gave an holistic account of what acquisition of knowledge and growth of truth mean in human history, on the one hand informed by the history of science to date and on the other hand anticipating future philosophy of science from Neurath to Quine.[53] But this story would have not found for a longtime any French philosopher able to tell it.

3 The Italian "Little Band"

Contemporaneous with James's philosophical adventures in Paris, a little group of young Italian intellectuals founded in Florence an audacious journal aiming to be the forum of anti-positivistic and anti-academic tendencies. Giovanni Papini and Giuseppe Prezzolini were the editors of *Leonardo*, which lasted from 1903 to 1907 – a short time which, however, was long enough for Papini and Prezzolini to become the two *enfants terribles* of Italian culture at the beginning of the twentieth century and, furthermore, for them to assume a leading role in the early diffusion of pragmatism in Italy.[54] James was highly sympathetic to the pragmatist philosophy arising in Italy and looked with admiration – as he stated in "G. Papini and the Pragmatist Movement in Italy", published on June 1906 in the *Journal of Philosophy* – at the "poor little Italy" where an "aggressive movement in favor of 'pragmatism'" was being born.[55] James appreciated the contributions written by "Signor Papini" and the "lightness, clearness and brevity" that the Florentine review offered in their monthly publications. James was in particular attracted by

[49](Berthelot 1911).

[50]*Id.*, p. 5.

[51]Regarding the failed reception of pragmatism as philosophy of science in France see (Brenner 2011, pp. 57–68).

[52](James 1975a, p. 93).

[53]On James and Quine see (Nevo 1995).

[54]See (Casini 2002). An excellent overview of pragmatism is still available in (Santucci 1963). For a very useful collection of essays on this topic see also (Maddalena and Tuzet 2007).

[55](James 1978, p. 144).

Papini's *Crepuscolo dei filosofi*, a book announcing – according to him – "the most radical conceiver of pragmatism to be found anywhere".[56] Moreover James was in Rome in the occasion of the 5th International Congress of Psychology (26–30 April, 1905) and he met – as he wrote immediately to his wife – "the little band" of Italian pragmatists, that is Papini, Prezzolini, Giovanni Vailati, Mario Calderoni, and Giovanni Amendola, having discussions with them. They show – he added in this letter – "an enthusiasm, and also a literary swing and activity that I know nothing of in our land, and that probably our dammed academic technics and PhD. machinery and university organization prevents from ever coming to a birth".[57] During the Congress in Rome Papini, Vailati and Calderoni gave three lectures on a common topic, that is "Beliefs and Will"(*Credenze e volontà*), collected than as "jumble limbs" in the issue of *Leonardo* appeared on June-August 1905.[58] It seems highly plausible that James embraced with great interest these contributions of the Italian Pragmatists to the Roman Congress. Moreover, at this time James was already profoundly attracted to Papini's philosophical way of reading the leading ideas of Pragmatism. It may be added that contrary to a widely accepted opinion, while in this lecture Papini endorsed free will and the pivotal function of beliefs, he did not yet endorse the kind of radical "voluntarism" that many interpreters have suggested was a typical feature of his thought.[59]

James was surely a great inspiration for the Florentine journal and he praised the Italian pragmatist club as "an extraordinarily free and spirited and unpedantic, group of writers".[60] But the pragmatism of his Italian admirers had two faces. As Papini remembered in 1913 in a collection of his pragmatist essays, *Leonardo* had endorsed two kinds of pragmatism: on the one hand a logical pragmatism, inspired by Peirce and developed by Vailati and Calderoni; on the other hand a "magic" pragmatism, supported by himself together with Prezzolini and totally devoted to stress action and free will.[61] Papini, in particular, was engaged in elaborating pragmatism not only from the perspective opened by the "will to believe", but also in the sense – so to speak – of approximating a divine will, aiming to dominate reality and to assimilate it through human mind. James and Nietzsche, the will to believe and *der Wille zur Macht* had to be coupled, according to Papini, in a strange marriage. In a less emphatic way Papini said: "We see the pragmatist kindled by a certain spirit of enthusiasm for all that shows the complexity and multiplicity of things, for whatever increases our power to act upon the world, for all that is most closely up with practice, activity, life".[62] James judged with veritable enthusiasm Papini's own interpretation of pragmatism. "What a thing is a genius!" –

[56] *Id.*, p. 145.

[57] See (Skrupskelis and Berkeley 2003, p. 26).

[58] The three lectures are available in (Frigessi 1960, pp. 270–279).

[59] Regarding this aspect see (Maddalena and Tuzet 2007, p. 163).

[60] (James 1978, p. 146).

[61] (Papini 1920, p. 8).

[62] *Id.*, p. 79.

he wrote to Papini in a letter of April 27, 1906 – "and you are a real genius!": so a great genius that James had no hesitation in calling him "the master of the movement now".[63] To some extent, one can say that Papini and the "little band" of *Leonardo* were themselves – in this crucial period – great inspirers of James's point of view. Not accidentally, in his famous *Pragmatism*, published about one year later, James recognized Papini's undeniable contributions to the development of the pragmatist movement and quoted the good metaphor Papini himself had used in his essay "Il pragmatismo messo in ordine" (1905). Papini conceived pragmatism as a corridor with many doors opening out from it and showing here a man on his knees invoking a religious faith and there another man engaged in eliminating metaphysics or in refashioning another one: by contrast, pragmatists themselves were to be characterized, according to Papini, as heroes passing through the corridor and looking at the future within a totally new perspective.[64]

But this is only a chapter of the whole story. The other face of Italian pragmatism at the very beginnings of twentieth century was the logical pragmatism developed by Vailati and his friend Calderoni. Vailati was undoubtedly one of the prominent intellectual figures of that time and found in Calderoni the best supporter of his own way to understand the pragmatic point of view.[65] Both Vailati and Calderoni were influenced by Peirce and his pragmatic maxim. Calderoni even published, in February 1905 in the journal *Leonardo*, an article opposing Papini and Prezzolini, suggesting that only the pragmatism of Peirce (understood as a sequel to the genuine positivism of John Stuart Mill) was able to succeed in criticising traditional philosophy, rejecting metaphysics, and applying a kind of experimental philosophy in the various field of human reasoning. Calderoni was convinced, on the one hand, that pragmatism could not be identified with the Jamesian "Gospel of the will to believe" professed by Papini and Prezzolini; nevertheless he recognized, on the other hand, that once belief was distinguished from the will – that is our beliefs as rules devoted to verification of propositions from the will to believe in the sense of the consequences for life derived from some presuppositions (e.g. God exists) – it was still plausible to consider positivism as a kind of pragmatism *à la* Peirce. The rules to verify meanings are simply rules formulating forecasts about what *could* be, starting from some premises expressed by linguistic propositions or practical intentions.[66]

For his part, Vailati was essentially in agreement with Calderoni. According to the received and in part still valid view, Vailati's conception of pragmatism was

[63](Perry 1936, vol. II, pp. 571–572).

[64](Papini 1920, p. 37). See also (James 1975a, p. 32): "As the young Italian pragmatist Papini has well said, [Pragmatism] lies in the midst of our theories, like a corridor in a hotel". And he adds: "But they all [i.e. the men in the chambers] own the corridor, and all must pass through it if they want a practicable onto or out of their respective rooms".

[65]On Vailati's Life and work see (De Zan 2000) and (De Zan 2009), also (Ferrari 2006, pp. 140–204).

[66](Calderoni 1924, vol. I, pp. 329–258).

quite different from Papini's and Prezzolini's. For Vailati, pragmatism possessed the significance of being able to account for scientific thought. So Vailati argued in his article published in *Leonardo* on February 1906 and entitled "Pragmatism and mathematical logic":

> One point of contact between logic and pragmatism is found in their common tendency to regard the value, and even the meaning of every assertion as being intimately related to the use which can be made, or which it may be desired to make, of it for the deduction and construction of particular consequences or groups of consequences.[67]

The main idea Vailati attempted to make plausible was that pragmatism – whose "initiator" was Peirce who promoted "an original trend in logical-mathematical studies" – was founded on a full convergence with mathematics and mathematical logic.[68] "Pragmatists and mathematicians", he wrote, "find themselves in agreement … in their efforts toward the maximum of conciseness and rapidity of expression – in their tendency to eliminate all superfluity and redundancy of wording and concept".[69]

For Vailati's understanding pragmatism Peirce was central. He regarded methodological rule of significance as fundamental for the new perspective opened by pragmatic attitude of thought.[70]

> Such methodological rule is nothing more than an invitation to translate our assertions into a form that makes it possible to apply, in an easier and more direct fashion, those very criteria of true and false which are more 'objective', less dependent on individual impressions an preferences. This form would be able to indicate more clearly what a kind of experiments or observations can and need to be performed, by us or others, to decide whether, and to what extent, our assertions are true.[71]

It would be nevertheless a misunderstanding to speak of a strict alternative between Peirce and James in Vailati's own interpretation of pragmatism. I would like to suggest that this point is a very important one in order to go beyond the standard view of Vailati as supporter of a "logical" type of pragmatism nourished by Peirce and contrasting James's version of the new way of thinking, which inspired the "magic" pragmatism endorsed by the Florentine Club. First and foremost we have to emphasize that some aspects of James's pragmatism are quite close to Vailati's own way of conceiving science, scientific experience, and knowledge. So Vailati and James agree in the refusal of a foundationalist epistemology and they agree on the essential role of idealisation in science; furthermore, both Vailati and James reject anything like "hard data" of experience and, more generally, any passive conception of experience. Last but not least, the fallibilistic conception of knowledge is a main feature common to Vailati and James, and Vailati – thanks also to his former

[67] (Vailati 2010, p.164).

[68] *Id.*, p. 163.

[69] *Id.*, p. 169.

[70] (Vailati 1911, p. 755).

[71] (Vailati 2010, p. 234).

collaboration to Giuseppe Peano's *Formulario di matematica* – shows very clearly the consequences of such a view in his reflections on the nature of definitions and postulates.[72]

In this context, Vailati's intellectual relationship with James offers indeed a wide range of important affinities, specifically with regard to philosophy of science and to some issues related to the field of scientific knowledge. The agreement between Vailati and James concerning an image of science quite different from the positivistic one can shed new light on the history of philosophy of science at the beginning of the twentieth century (from Mach to Duhem) and, at once same time, can contribute to a revaluation of James's epistemology beyond the received view. To begin with, although Vailati had undoubtedly a great admiration for Peirce's pragmatic rule of meaning (i.e. the rule formulated by Peirce in his seminal essay *How to Make Clear our Ideas*), he was also immediately aware of the *epistemological* relevance of James's Pragmatism. In his too often ignored reviews both of the *Will to believe* and, some years later, of James's most famous *Pragmatism*, Vailati emphasized the great merit of James's rehabilitation of "the *constructive* and *anticipating* activities of human understanding". According to Vailati, James was right in criticizing the usual conception of scientific and philosophical truth maintained by positivism, which had underestimated such a view of knowledge and consequently has endorsed an image of mental activity which is limited to a mere classification and, so to speak, to a recording of empirical data. In Vailati's opinion, James is in this sense totally in agreement with the recent "logic of science", namely with the analysis developed by Mach, Clifford and others of methods, history and principles of modern science. On the other hand, Vailati underlined the epistemological importance of James' critical assessment of positivism as well as of the sometimes "narrow-minded" philosophy nourished by the scientists. Vailati totally agreed with James's emphasis on the crucial role in the scientific inquiry of audacious formulation of hypotheses[73]; similarly, he pointed out that James recognized better than any other philosopher of science the function of belief for the scientific method.[74] Broadly speaking, Vailati appreciated James's view that scientific knowledge is always the result of a *mental construction* where the empirical, factual basis was not as incorrigible as the (positivistic) standard view was inclined to suggest.[75]

Vailati's great merit was to have understood, quite differently from some of his contemporaries in France, that James was elaborating a version of pragmatism that was in no way to be thought of as a merely "irrationalistic" philosophy. For his part James was fully convinced of the need for an anti-foundationalist account of knowledge by realizing that our ideas develop in a very different way from that typically described by traditional philosophy since Descartes. In just this sense

[72](Vailati 1911, pp. 449–453, 761–764).

[73]*Id.*, p. 270.

[74]*Id.*, p. 270.

[75]*Id.*, p. 283.

James gave an original account of recent developments in science in "Humanism and Truth", an article from which Vailati also quoted.

> Up to about 1850 almost everyone believed that sciences expressed truths that were exact copies of a definite code of non-human realities. But the enormously rapid multiplication of theories in these latter days has well-nigh upset the notion of any one of them being a more literally objective kind of things than another. There are so many geometries, so many logics, so many physical and chemical hypotheses, so many classifications, each one of them good for so much and yet not good for everything, that the notion that even the truest formula may be a human device and not a literal transcript has dawned upon us. We hear scientific laws now treated as so much 'conceptual shorthand', true so far as they are useful but not farther. Our mind has become tolerant of symbol instead of reproduction, of approximation instead of exactness, of plasticity instead of rigor.[76]

It is also proper to state that Vailati was right in emphasizing the epistemological core of James' pragmatism and that in this he proved a laudable exception in the philosophical landscape at the beginning of twentieth century in Europe. In this Vailati agreed with an outsider such as the Viennese *Privatdozent* Wilhem Jerusalem, the translator into German of James's *Pragmatism* and who was, in his turn, a very interesting exception within the German speaking philosophical community, which was instead resolved to reject the "Yankee" philosophy just arrived in Europe. It is noteworthy to remember, that Vailati met Jerusalem on the occasion of the 3rd International Congress of Philosophy in Heidelberg and that starting from just this event another story about the diffusion of pragmatism in Europe could be told – a story which has to do with James, Mach, Vailati, Jerusalem, the origins of logical empiricism and the emigration of ideas not only from the Vienna Circle to Harvard Square, but also from Harvard Square to Vienna Circle.[77] But this is another story, and it would be necessary to leave French and Italy if we would like to describe this new adventure of thought.

References

Bergson, H. 2001. *Écrits philosophiques*. Paris: Puf.
Bergson, H. 2011. *Sur le pragmatisme de William James*. Paris: Puf.
Berthelot, R. 1911. *Un romantisme utilitaire. Étude sur le mouvement Pragmatiste, vol. I, Le Pragmatisme chez Nietzsche et chez Poincaré*. Paris: Alcan.
Berthelot, R., et al. 1908. La signification du pragmatisme. *Bulletin de la Société Française de Philosophie* 8: 249–296.
Boutroux, É. 1908. *Science et Religion dans la philosophie contemporaine*. Paris: Flammarion.
Boutroux, É. 1911. *William James*. Paris: Colin.
Brenner, A. 2011. Bergson, James et le pragmatisme scientifique. In *Bergson et James. Cent ans après*, ed. S. Madelrieux, 57–68. Paris: Puf.
Calderoni, M. 1924. *Scritti*, 2 vol., ed. by O. Campa. Firenze: La Voce.

[76](James 1975b, p. 206).

[77]On this topic I allow myself to refer to (Ferrari 2010a) and (Ferrari 2010b). See also (Holton 1993).

Carlson, Th. 1997. James and the Kantian tradition. In *The Cambridge companion to William James*, ed. R.A. Putnam, 363–383. Cambridge: Cambridge University Press.
Casini, P. 2002. *Alle origini del Novecento. «Leonardo», 1903–1907*. Bologna: Il Mulino.
Couturat, L. 1983. Leçon inaugurale au Collège de France 8 dicembre 1905. In *L'oeuvre de Louis Couturat (1868–1914) . . . de Leibniz à Russell . . .*, 17–33. Paris: Presses de l'Ecole Normale Supérieure.
De Zan, M. (ed.). 2000. *I mondi di carta di Giovanni Vailati*. Milano: Franco Angeli.
De Zan, M. 2009. *La formazione di Giovanni Vailati*. Galatina: Congedo Editore.
Ferrari, M. 2006. *Non solo idealismo. Filosofi e filosofie in Italia tra Ottocento e Novecento*. Firenze: Le Lettere.
Ferrari, M. 2010a. William James a Vienna. *Paradigmi* 28: 97–115.
Ferrari, M. 2010b. Heidelberg 1908. Giovanni Vailati, Wilhelm Jerusalem e il pragmatismo Americano. *Giornale critico della filosofia italiana* 89: 9–31.
Frigessi, D. (ed.). 1960. *La cultura italiana del '900 attraverso le riviste,* vol. I, *«Leonardo», «Hermes», «Il Regno»*. Turin: Einaudi.
Hill, H. 2009. Pragmatisme en France. The case of Édouard Le Roy. In *The reception of pragmatism in France & the rise of Roman Catholic modernism, 1890–1914*, ed. D.G. Schultenover, S.J., 143–166. Washington, DC: The Catholic University of America Press.
Holton, G. 1993. From the Vienna circle to Harvard square: The Americanization of a European world conception. In *Scientific philosophy: Origins and developments*, ed. F. Stadler, 47–73. Dordrecht/Boston/London: Kluwer.
James, W. 1975a. *Pragmatism. A new name for some old ways of thinking*. Cambridge, MA: Harvard University Press.
James, W. 1975b. *The meaning of truth. A sequel to pragmatism*. Cambridge, MA: Harvard University Press.
James, W. 1978. *Essays in philosophy*. Cambridge, MA/London: Harvard University Press.
James, W. 1979. *Some problems of philosophy*. Cambridge, MA/London: Harvard University Press.
James, W. 1996. *A pluralistic universe*. Lincoln/London: University of Nebraska Press.
Le Roy, É. 1913. *Une philosophie nouvelle. Henri Bergson*. Paris: Alcan.
Maddalena, G., and G. Tuzet (eds.). 2007. *I pragmatisti italiani. Tra alleati e nemici*. Milano: Albo Versorio.
Madelrieux, S. (ed.). 2011. *Bergson et James. Cent ans après*. Paris: Puf.
Nevo, I. 1995. James, Quine, and analytic pragmatism. In *Pragmatism. From progressivism to postmodernism*, ed. D. Hollinger and D. Depew, 153–161. Westport/London: Prager.
Papini, G. 1920. *Pragmatismo (1903–1911)*. Firenze: Vallecchi.
Parodi, D. 1919. *La philosophie contemporaine en France. Essai de classification des doctrines*. Paris: Alcan.
Peirce, C.S. 1879. Comment rendre nos idées claires. *Revue Philosophique de la France et de l' Étranger* 6: 553–569.
Perry, R.B. (ed.). 1929. Correspondence de Charles Renouvier et de William James. *Revue de métaphysique et de Morale* 36: 1–35, 193–222.
Perry, R.B. (ed.). 1935. Correspondence de Charles Renouvier et de William James. *Revue de métaphysique et de Morale* 57: 303–318.
Perry, R.B. 1936. *The thought and character of William James*, 2 vols. Boston: Little, Brown, and Company.
Pillon, F. 1867. Avvertissement. *L'Année Philosophique* 1: V–VI.
Ranchetti, M. 1963. *Cultura e riforma religiosa nella storia del modernismo*. Turin: Einaudi.
Renouvier, C. 1859. *Essais de critique générale, vol. II, L'homme: la raion, la passion, la liberté. La certitude, la probabilité morale*. Vrin (first edition 1929).
Renouvier, C. 1867. Introduction. De la philosophie du XIXe siècle en France. *L'Année Philosophique* 1: 1–108.
Santucci, A. 1963. *Il pragmatismo in Italia*. Bologna: Il Mulino.
Scoppola, P. 1961. *Crisi modernista e rinnovamento cattolico in Italia*. Bologna: Il Mulino.

Shook, J.R. 2009. Early responses to American pragmatism in France. Selective attention and critical reaction. In *The reception of pragmatism in France & the rise of Roman Catholic modernism, 1890–1914*, ed. D.G. Schultenover, S.J., 59–75. Washington, DC: The Catholic University of America Press.

Skrupskelis, I.K., and E.M. Berkeley (eds.). 2003. *The correspondence of William James, volume 11, April 1905–March 1908*. Charlottesville/London: University of Virginia Press.

Vailati, G. 1911. *Scritti*, ed. M. Calderoni, U. Ricci, and G. Vacca. Firenze/Leipzig: Seeber & Barth.

Vailati, G. 2010. *Logic and pragmatism. Selected essays*, ed. C. Arrighi, P. Cantù, M. De Zan, and P. Suppes. Stanford: CSLI Publications.

Wahl, J. 2004. William James d'après sa correspondance. In *Vers le concret. Études d'historie de la philosophie contemporaine*, 47–117. Paris: Vrin.

Worms, F. 1999. Bergson et James. Lectures croisées. *Philosophie* 64: 54–68.

European Pragmatism? Further Thoughts on the German and Austrian Reception of American Pragmatism

Thomas Uebel

1 Introduction

Massimo Ferrari has provided us with a masterful survey of the French and Italian reception of American pragmatism in the early twentieth century, with a distinctive focus on the philosophy of science.[1] Once more we see in action the great variety of philosophical voices that populated Europe a century ago: in this context, we see very positive reactions to the new philosophy from the "new world". To extend the survey a little bit I want to look at the reception of American pragmatism elsewhere in Europe.[2] In particular I want to return to the discussion of the reception of pragmatism in Germany and Austria begun by Ferrari at our network's first plenary conference and like him I want to focus on philosophy of science.[3] As is well-known, pragmatism's reception there was rather hostile, with only very few exceptions. I want to look at the major one of these to see whether we can derive from it any hints as to what accounts for the exceptions as well as the far more common hostility. Here I cannot, however, defend but only develop a thesis that I hope may be assessed more fully on another occasion.

[1] Ferrari (2014).

[2] On the reception in England and particularly the role of Ferdinand Canning Scott Schiller (and some brief remarks about the reception elsewhere), see Shook (2004).

[3] See Ferrari (2010).

T. Uebel (✉)
Philosophy, School of Social Sciences, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
e-mail: thomas.uebel@manchester.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_44,

2 The Austro-German Reception of American Pragmatism

When talking about the reception of American pragmatism in Germany and Austria before World War I, our focus is mainly William James and his book *Pragmatism. A New Name for some Old Ways of Thinking*. Published in America in early 1907, its German translation by the Viennese pedagogue and philosopher Wilhelm Jerusalem was out already at the end of the same year (but carried the imprint 1908).[4] Prior to this, translations of James's essay collection *The Will to Believe* (orig. 1887) had been published in 1899, of his *Talks to Teachers on Psychology* (orig. 1899) in 1900 and of his *Varieties of Religious Experience* (orig. 1902) earlier in 1907; after *Pragmatismus*, there followed translations of *Psychology. Briefer Course* (orig. 1892) in 1909 and of *The Pluralistic Universe* (orig.1909) in 1914.[5] (In between the latter two, also a selection of translations of essays from F.C.S. Schiller's *Humanism. Philosophical Essays* (orig. 1903) and *Studies in Humanism* (orig. 1907) was published in 1911.[6]) So James was certainly known and, as a psychologist, respected. However, apart from Jerusalem's early championship of James's pragmatism – most prominently so at the International Congress for Philosophy in Heidelberg in September 1908 and in articles in *Deutsche Literaturzeitung*[7] – the Central European philosophy professoriat and those who aspired to ascend to its ranks rejected pragmatism as problematical if not deeply unphilosophical and superficial.[8] (Günther Jacoby's initial sympathy soon dissipated in his subsequent defense of the German metaphysical intellect against James's "assault" on it.)[9] Importantly, apart from rare cognoscenti like Ludwig Stein, the founder of the *Archiv für Geschichte der Philosophie*, the work of Charles Sanders Peirce remained unknown.[10]

The situation did not change much after World War I when, again, no sympathetic discussion of the pragmatist movement and its aims by professional German or Austrian philosophers – though plenty of unsympathetic ones even by prominent exponents of the field as far apart as Max Scheler and Max Horkheimer[11] – can be found until the late 1930s.[12] Peirce remained largely unknown until his *Collected Papers* begun to be published in 1931 and reviewed, first in *Kant-Studien* in 1933

[4] James (1907a, 1908).

[5] James (1899, 1900, 1907b, 1909, 1914).

[6] Schiller (1911).

[7] See Jerusalem (1908a, b, 1909, 1910a, 1913). Another exception is (Vorbrodt 1913), a sympathetic exposition of the content of a monograph by the Swiss psychologist (Flournoy 1911).

[8] For accounts of the mostly hostile reception, see Oehler (1977), Dahms (1992), and Ferrari (2005).

[9] See Jacoby (1909, 1912a, b).

[10] See the thoughtful assessment of the challenges facing pragmatism in Stein (1908).

[11] For references see the papers listed in Fn. 8 above.

[12] By contrast, in newly independent post-World War I Czechoslovakia philosophy saw a fairly intense reception of pragmatism which included Peirce; see Capek (1918) and Vorovka (1929); cf. (Shook 2004, pp. 52–53).

and then in *Deutsche Literaturzeitung* by the Munster logician Heinrich Scholz in 1934 and 1936.[13] A wider reception of Peirce had to wait until after World War II. Finally, to be sure, translations of a few works by Dewey were published in German in the early 1930s (*Democracy and Education*, orig. 1916, and *Human Nature and Conduct*, orig. 1922, and some pedagogical essays),[14] but as with James, being available in translation as a psychologist and educator did not translate into a sympathetic reception amongst academic philosophers. The sole exception was Eduard Baumgarten's second volume of his study of American thought *Die geistigen Grundlagen des amerikanischen Gemeinwesens* published in 1938 and favourably reviewed still in the same year in *Journal of Philosophy*.[15] Unlike Arnold Gehlen's philosophical anthopology *Der Mensch. Seine Natur und seine Stellung in der Welt* of 1941, which endorsed pragmatism in rather more general terms, Baumgarten's book was kept largely free of tributes to the then dominant German ideology.[16]

Among post-World War I philosophers of science – who did not necessarily occupy chairs of philosophy – the situation appears to have been the same, with one notable exception: among the members of the Vienna Circle was one, Philipp Frank, who celebrated the pragmatist tendencies of early logical empiricism already in 1929, years before exile in the United States made such talk more generally expedient.[17] (Frank focussed mainly on James's anti-correspondentism.) But even in the Circle Frank met with opposition, namely from Schlick who had opposed the pragmatist theory of truth ever since 1910. But thereby hangs another complicated tale; I will now focus on the period prior to World War I.[18]

The only academic philosopher of any consequence then – and this need not mean of terribly great consequence – who we can point to in pre-World War I Central Europe as having adopted pragmatist thought of the American variety, explicitly under this heading, was Jerusalem.[19] Jerusalem was not really a philosopher of science, but having come from pedagogics and psychology he was not a typical German or Austrian philosopher either.[20] In fact, given his admiration for and friendship with Ernst Mach, his own naturalistic methodology and his

[13]See Metz (1933) and Scholz (1934, 1936); see also Müller (1931).

[14]Dewey (1930, 1931) and Dewey and Kilpatrick (1935).

[15]See Baumgarten (1938) and Schneider (1938). The review is signed only "H.W.S" which, given the authority with which it is written I presume to stand for Herbert Wallace Schneider; on Schneider see Walton and Anton (1974).

[16]Gehlen (1941); for the comparison see Dahms (1987).

[17]Frank (1929).

[18]See now Uebel (2014).

[19]There was also Hans Kleinpeter, who gave a sympathetic reception to pragmatism from a Machian perspective in Kleinpeter (1911/1912) and noted Nietzsche as a precursor of pragmatism in Kleinpeter (1913). But as a high school teacher who, unlike Jerusalem, never received a call to teach at university, Kleinpeter's voice carried less weight.

[20]On Jerusalem see Eckstein (1935); for a recent assessment see Uebel (2012).

increasing interest in the sociology of cognition, he came perhaps as a close to being a philosopher of science as it was possible without being one. As we shall see, this is not insignificant. What sympathetic reception pragmatism had in contemporary German philosophical thought was largely limited to philosophy of science – one rogue Kantian excepted – and there mainly to practicing philosopher-scientists or thinkers allied to them. This makes for an interesting contrast with France, even Italy, as Ferrari described it. James's type of metaphysics found little reception – not, however, because German philosophy was wholly rationalistic, but because its own *Lebensphilosophie* also felt the need to distinguish itself from perceived utilitarianism of Anglo-American pragmatism.[21]

3 Wilhelm Jerusalem's Advocacy of Pragmatism

What is noteworthy is Jerusalem's own "conversion" to pragmatism. According to his own account, his prior work played a crucial role, in particular *Der kritische Idealismus und die reine Logik* of 1905, a strongly polemical intervention in the psychologism-dispute then raging in, even dominating German philosophy. (This was a wide-ranging *Streit* which concerned both the specific doctrine that the laws of logic and arithmetic are ultimately of an empirical nature and the more general issue of the proper method of philosophical inquiry; in addition it provided the occasion for the final disciplinary separation of psychology from philosophy.)[22] It was precisely his opposition to a priori philosophising that prompted the English pragmatist F.C.S. Schiller in a review of Jerusalem's 1905 book to comment on his proximity to pragmatism.[23] This in turn led Jerusalem to inquire about this movement with James – with whom he already had been in correspondence about psychological matters – and which ultimately led to Jerusalem becoming James's translator.[24] This path "into" pragmatism strikes me as highly significant. It highlights an important aspect of what kind of doctrine pragmatism was thought to be when it was accepted by Jerusalem and rejected by most other German philosophers.

Consider Jerusalem's pronouncements on truth that had attracted Schiller's attention:

> Truth ... is created only by the function of judgement. ... In judgement human beings of the most primitive level of development only adopt a stance, however. This adoption of a stance consists in the actions which are prompted by the interpretation of a perceived process. If the measures taken on the basis of that interpretation prove to be beneficial for life, biologically useful, then the interpretation was right; if they prove to be superfluous

[21] On Lebensphilosophie generally see, e.g., Schnädelbach (1984, ch. 4).

[22] See Kusch (1995).

[23] Schiller (1906).

[24] See Jerusalem (1925, pp. 32–33).

> or even detrimental then the interpretation was wrong. … The valuation which action is accorded due to the benefit or detriment it brings with it, this valuation and nothing else is the origin of the concepts *true* and *false*. (1905, p. 162, trans. TU)

So far, so (proto-) pragmatic and anti-Platonistic: no judgement, no truth.[25] Jerusalem went on:

> Soon the function of judgement proves so valuable and beneficial for life that it is exercised even where there is no immediate employment of its interpretation. We pass judgements in advance for later, as it were, and store the results of the interpretations made in our memory. … This brings with it a change in the meaning of the concept of truth. A judgement is true in the first stage of development only in so far as it prompts us immediately to undertake measures that are useful and beneficial for life. As soon as we begin to judge for storage, however, this meaning becomes a broader one. … Purely theoretical, highly objective judgement is thus itself a product of the instinct of preservation, indeed one of its most valuable and significant products. … Such judgements are in fact possible – as proved by the emergence and development of science. There research into the laws governing physical and psychological processes is undertaken without concern for practical gain. Even if the results of these investigations in the end are destined to make the life of humanity safer, richer and more pleasurable, the individual researcher can contribute to this last and highest aim only by proceeding strictly objectively. (*Id.*, pp. 168–169, trans. TU)

Given the kind of opposition to pragmatism that he was to run into later, it is notable that Jerusalem here did not deny "purely theoretical" truth (what he later was to call "objective truth"): "It is possible for us to think theoretically, but first we had to learn to do so." (*Id.*, p. 170) We may note then that Jerusalem, with his own form of philosophical naturalism, anticipated the pragmatists' rejection of correspondence truth as a philosophical primitive as much as their rejection of a priori theorising and its replacement by reflections on human cognitive capacities in their evolutionary role. Just those were the ideas, after all, that Jerusalem foregrounded in the pragmatism he defended at Heidelberg and ever since. As he stressed in his "Translator's Preface": pragmatism was primarily a method of inquiry, not a system of philosophical propositions (1908c, p. iv).

What the quotations from the pre-pragmatist Jerusalem make evident is that when he did become a pragmatist, he really did not have to change his views at all – all he had to do was re-label them and advertise, as he did, his own evolutionary-historical elaborations of their theory of truth and its embedding in the overall socio-cultural development of mankind as "supplementations" (*id.*, p. vi). Accordingly, when in presenting pragmatism he asserted that "[f]or pragmatists, there is no such thing as a purely theoretical truth the content of which would never be practically discernible anywhere" (1908a, p. 138), the critical emphasis lay on the concept of a *discernible* difference: Jerusalem endorsed what James called "Peirce's principle" (on which more below).

[25] In these passages Jerusalem also called upon similar ideas in Simmel (1900, pp. 58–66), which built on ideas first expressed in Simmel (1895). In later years, however, Simmel was critical of pragmatism: see Ferrari (2005).

4 The Specificity of the Austro-German Context

This "anticipation" of pragmatism's leading ideas of anti-correspondentism and philosophical naturalism by Jerusalem gives a strong hint as to the nature of the "reception" of pragmatism by those then contemporary German and Austrian philosopher-scientists who did prove receptive to it. It is that in so far as they did confess pragmatist ideas, they in fact confessed to ideas that they themselves (or leading colleagues) had developed in roughly the same time period but had developed independently of the American pragmatists.[26] Their "pragmatism", in other words, was *homegrown*.

It is this thesis that I want to motivate here. Were it correct it would help to explain the reactions of German and Austrian philosophers when confronted with American pragmatism from 1908 onwards. To do so, of course, we must also place the issue of their reception of pragmatism in its local context and in so doing progress beyond invoking generalized talk of the German mentality or its typical philosophical orientation, however much that may be a contributing factor.[27] It is not merely and certainly not always the case that the early opponents of pragmatism were hopelessly enthralled by idealist metaphysics.[28] Rather German and Austrian philosophers adopted their stance towards American pragmatism for either of two reasons. Either they had opposed or they had championed broadly related views for a considerable time already – albeit under different and varied names that refer back to the above mentioned long-standing debate about psychologism.

Consider Jerusalem. In his talk at the Heidelberg congress defending pragmatism we read: "Even the most universal propositions of logic and mathematics are regarded" – by the evolutionist in contrast to the apriorist thinker – "only as sedimentations, as condensations of earlier experience. The evolutionist sees in these propositions the adaptation of thoughts to facts and to each other (Mach), he finds in these valuable measures from the point of the economy of thought" (1909, p. 809). In *Pragmatism* James only claimed that "the form and order" of "those bodies of truth known as logics, geometries, or arithmetics" is "flagrantly man-made" and that "mathematics and logic themselves are fermenting with human rearrangements" (1907a/1991, pp. 108 and 112). This is not decisive, but since Jerusalem argued along similar lines to press his psychologistic conclusions, and James gave no grounds to argue against these, it was not an unreasonable conclusion to associate pragmatism and psychologism.[29] Since then due to its

[26] "In the new pragmatic method ... I have found a theory to which I have been led independently by my own investigations even before many of its American exponents" (Jerusalem 1910b, p. vi).

[27] Shook overshoots the mark in claiming that "unlike France or England, Germany had no ongoing native movement struggling against rationalism" – be that *Lebensphilosophie* or, considerably more soberly, Jerusalem in Austria; see Shook (2004, pp. 51–52).

[28] Consider the logician Ernst Mally's response to Jerusalem's Heidelberg congress talk in Elsenhans (1909, p. 814).

[29] See also the clear endorsement of psychologism in Schiller (1907, p. xii).

Austrian protagonist's views it was easily characterised as a philosophy entailing a version of psychologism, most German and Austrian academic philosophers were naturally opposed to pragmatism.[30] On the other hand, those few philosophers and scientists who had come to adopt an overall evolutionist and anti-metaphysical perspective on matters of mind and knowledge were able to find themselves at least in partial sympathy with pragmatism (even if for various reasons they did not consider themselves as pragmatists).

To consider further the thesis that what Austro-German support there was for American pragmatism was largely home-grown in that it built on views arrived at independently in that very specific intellectual context, let us take a brief look at the "positivism" of Mach. Mach too was no stranger to psychologism.[31] That alone, of course, would not account for much of a theoretical or methodological convergence of his views with pragmatism (nor did that alone do so in the case of Jerusalem). So what else did Mach bring to the table he (partly) shared with James, as it were?

5 The Pragmatism in Mach's Positivism

That his personal acquaintance with Mach on the occasion of his visit to Prague in 1882 left a lasting impression of his "pure intellectual genius" on James is well-known.[32] But his first biographer also wrote:

> From Mach, James had learned something of what he knew about the history of science, and he had readily accepted his view of the biological and economic function of scientific concepts. That was in the early days. In his insistence on the practical motives of knowledge, James had now [ca. 1907, TU] gone so far that he could obtain from Mach only the sense of a somewhat lagging support: 'W.J., but not emphatic enough,' he wrote in the margin of his copy of Mach's *Erkenntnis und Irrtum*. The latter's acknowledgement of *Pragmatism* was brief and somewhat perfunctory: 'I have read the book through hurriedly from beginning to end with great interest, in order to study it again later. Although I am by my training a scientist and not at all a philosopher, nevertheless I stand very close to pragmatism in my way of thinking, without ever having used that name. By following out his line the unprofitable differences between philosophers and scientists could be resolved.' (Perry 1936, p. 463)

[30] See Kusch (1995) for an exhaustive inventory of anti-psychologistic positions taken; pragmatism does not figure in his discussion.

[31] E.g.: "The greatest perfection of mental economy is attained in that science which has reached the highest formal development, and which is widely employed in physical inquiry, namely, in mathematics. … No one will dispute me when I say that the most elementary as well as the highest mathematics are economically ordered experiences of counting, put in forms ready for use" (Mach 1882/1943, p. 195). See also Mach (1896/1986, pp. 410–411).

[32] See the often-quoted letter to his wife, 2 November 1882, e.g. in Thiele (1978, p. 169).

Perry's judgement that Mach provided only lagging support needs to be qualified somewhat.[33] But what is true and important for our purposes is that Mach's support of James's pragmatism extended only to what pertained to the understanding of science. Mach did not care for James's metaphysical-theological flights of fancy as his correspondence – albeit not with James – shows very clearly.[34]

The same evaluative difference also found expression in Perry's overall correct summary of James's reception in Germany:

> Although Ernst Mach was an important forerunner of pragmatism, while Simmel and Ostwald were greeted by James as allies, pragmatism gained only a light foothold in Germany, and that mainly in Austria! Even these three philosophers just mentioned accepted it as an interpretation of method in the physical or social sciences rather than as a philosophy. (1936, pp. 579–580)[35]

Here I now want to ask: once we discount James's metaphysics, what remains of his pragmatism that has not already been developed by his "forerunner" Mach? To answer this question we need only consider the more philosophical chapters of Mach's main historical works, *The Science of Mechanics* of 1883 and related essays, *Principles of the Theory of Heat* of 1896, and, of course, his later *Knowledge and Error* (*Erkenntnis und Irrtum*) of 1905.[36]

Before we turn to these works, however, it is important to remember that the aim of all of Mach's histories was philosophical after all, in the sense of laying bare the epistemological principles in accordance with which the different branches of physics developed. Thus he remarked early on in *History and Root of the Principle of the Conservation of Energy*: "We are accustomed to call concepts metaphysical, if we have forgotten how we reached them." (1872/1911, p. 17) His remedy was plain: "There is only one way to enlightenment: historical studies." (*Id.*, p. 16) All along

---

[33]Perry also noted about Mach that "his last work approached closely to the pragmatist position" (Perry 1936, p. 588); presumably he meant *Erkenntnis und Irrtum* of 1905. Note also that in 1903 Mach had dedicated the third edition of his *Popular Scientific Lectures*, containing seven new essays, to William James and retained the dedication for the fourth edition in 1910.

[34]Mach to his Danish colleague Anton Thomsen, 21 January 1911, recalling their encounter in Prague and assessing James's philosophy: "I cannot think of anyone with whom I was able to discuss matters as well and as fruitfully despite the divergence in our views [as with him]. He opposed me nearly in everything and yet I gained in nearly everything from his objections. ... The main focus of his work lies certainly in his excellent psychology. I cannot quite agree with his pragmatism [nicht ganz befreunden]. 'We must not drop the concept of God because it promises us too much.' That is a dangerous argument. 'It is not only the healthy ones who have the correct insight .' There is truth in this, but it would be sad if the judgement of those who are healthy is directed by those who are not. The world must be intelligible primarily to those who are healthy" (Blackmore and Hentschel 1985, p. 86; trans. TU).

[35]A correction is needed here: James did not mention Simmel in *Pragmatism*, but Jerusalem did in his "Translator's Preface" (1908c, p. v); for the relevance of Simmel for Jerusalem see Fn. 25 above.

[36]See Mach (1882/1943, 1883/1960, ch. 4, Sect. 4, 1884/1943, 1896/1986, chs. 25–34, 1905/1976, passim).

Mach's methodology in these works was "historical-critical" (to use the subtitle of his *Science of Mechanics*).

Following his investigations along this path, Mach concluded about scientific theories in general that they operate according to a principle of mental economy:

> If all individual facts – all the individual phenomena knowledge of which we desire – were immediately accessible to us, a science would never have arisen. Because the mental power, the memory, of the individual is limited, the material must be arranged. . . . [a] 'law' has not in the least more real value than the aggregate of the individual facts. Its value for us lies merely in the convenience of its use: it has an economical value. (*Id.*, pp. 54–55)

In terms borrowed from his friend, the political economist Emanuel Herrmann, Mach believed that "science faces a problem of economy or thrift" (*id.*, p. 88, trans. altered).[37] Science operated within a practical context where it was constrained by what's thinkable and doable with the resources at hand. But as has often been remarked, economies can be effected in different ways. Mach was concerned therefore to stress that arriving by analysis at simpler elements must not be misunderstood as a different kind of achievement than it was.

> Besides this collection of as many facts as possible in a synoptical form, natural science has yet another problem which is also economical in nature. It has to resolve the more complicated facts into as few and as simple ones as possible. This we call explaining. These simplest facts, to which we reduce the more complicated ones, are always unintelligible in themselves, that is to say, they are not further resolvable. . . . Now it is only, on the one hand, an economical question, and, on the other, a question of taste, at what unintelligibilities we stop. People usually deceive themselves in thinking that they have reduced the unintelligible to the intelligible. Understanding consists in analysis alone; and people usually reduce uncommon unintelligibilities to common ones. They always get, finally, to . . . propositions which must follow from intuition and, therefore, are not further intelligible. (*Id.*, pp. 55–56)

Here Mach's well-known philosophical deflationism is plainly in evidence. Analysis provided understanding alright, but this understanding exhibited neither the depth nor the certainty and irrevisability which apriorist metaphysicians sought to endow their own intuitions with. He concluded:

> In the investigation of nature, we always and alone have to do with the finding of the best and the simplest rules for the derivation of the phenomena from one another. One fundamental fact is not at all more intelligible than another: the choice of fundamental facts is a matter of convenience, history and custom. (*Id.*, p. 57)

With this anti-realism Mach early on distinguished his own position from that of most contemporaries, including that of his later friend and admirer Jerusalem. (Unlike Mach, Jerusalem was also not wholly repelled by James's metaphysics even though he did not endorse it.) Mach urged: "Let us not let go of the guiding hand of history. History has made all; history can alter all." (*Id.*, p. 18)

Around the time that James visited him in Prague, some 10 years later, Mach had further broadened the horizon of his historical-critical inquiries and at the same time deepened them. As he put in *The Science of Mechanics*: "In the reproduction

[37] On Herrmann, see Haller (1986).

of facts in thought, we never reproduce the facts in full, but only that side of them which is important to us, moved to this directly or indirectly by a practical interest. Our reproductions are invariably abstractions." (1883/1960, pp. 578–579). It was the evolutionary origin of this interest-relativity that Mach began to stress. What moved him was not the cultural particularism that inspired some of his contemporaries to revive German idealism (and class and ethnic privilege) under the heading of "historicism". Rather he "found it helpful and restraining to look upon every-day thinking and science in general, as a biological and organic phenomenon, in which logical thinking assumed the position of an ideal limiting case" (*id*., p. 593).[38] To better understand the development of science and its epistemology, he now sought "to consider the growth of natural knowledge in the light of the theory of evolution. For knowledge, too, is a product of organic nature." (1884/1943, p. 217). Or, as he later summarised it in the opening sentence of *Knowledge and Error*: "Scientific thought arises out of ordinary thought, and so completes the continuous series of biological development that begins with the first simple manifestation of life." (1905/1976, p. 1, trans. altered).

Importantly, Mach differentiated his own evolutionism from the Social Darwinism peddled in his day and since:

> Art and science, any ideas of justice and ethics, indeed any higher intellectual culture can flourish only in the social community, only when one part relieves the other of some of its material cares. Let the 'upper ten thousand' recognize clearly what they owe the working people! Let artists and scientists reflect that it is a great common and jointly acquired human possession that they administer and extend! (*Id.*, p. 61)

With his political-ideological position clarified, we may now note that Mach discerned the so-called principle of the economy of thought in nearly all aspects of scientific inquiry and always traced it back to its ultimately evolutionary context.

> It is the object of science to replace, or *save*, experiences, by the reproduction and anticipation of facts in thought. Memory is handier than experience and often answers to the same purpose. This economical office of science, which fills its whole life, is apparent at first glance; and with its full recognition all mysticism in science disappears. Science is communicated by instruction, in order that one man may profit by the experience of another and be spared the trouble of accumulating it for himself; and thus to spare posterity, the experiences of whole generations are stored up in libraries. (1883/1960, p. 577, orig. emphasis)

Whatever the practical purposes of inquiry may be that set the parameters of convenience now, all cognition arose originally as means to survival. Being evolutionary in origin the principle of economy of thought extended to the material means of representation as much as to its content: "Language is itself an economical contrivance." (*Id.*, p. 578) In terms of content the principle could be traced throughout the development of science, both in specific doctrines and general features of theory

[38]This is from a passage added in the 4th edition (1901), but the evolutionary perspective is clearly discernible already in the first edition version of the section "The Economy of Science", to which this is an addition.

formation like the determination of laws of nature.[39] "In the details of science, its economical character is still more apparent." For instance, "in nature there is no law of refraction, only different cases of refraction. The law of refraction is a concise compendious rule, devised by us for the mental reconstruction of a fact . . ." (*Id.*, p. 582; cf. 1882/1943, p. 193; 1884/1943, p. 231; 1896/1986, p. 357) In sum: "Science itself . . . may be regarded as a minimal problem, consisting of the completest possible representation of facts with the least possible expenditure of thought." (1883/1960, p. 586)[40] James, who studied Mach's *Science of Mechanics* closely in the years after his visit to Prague,[41] undoubtedly made this reasoning his own.

Now note that as part of his general naturalistic approach Mach also formulated a maxim for scientific reasoning that has a good claim of being placed next to what James called "Peirce's principle".

> The function of science, as we take it, is to replace experience. Thus, on the one hand, science must remain in the province of experience, but, on the other, must hasten beyond it, constantly expecting confirmation, constantly expecting the reverse. *Where neither confirmation nor refutation is possible, science is not concerned.* (1883/1960, pp. 586–587, emphasis added)

Peirce's principle was: "Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of those effects is the whole of our conception of the object" (Peirce 1878/1992, p. 132). James's paraphrase of the principle was this:

> To attain perfect clearness in our thoughts of an object, then, we need only consider what conceivable effects of a practical kind the object may involve – what sensation we are to expect from it, and what reactions we must prepare. Our conception of these effects, whether immediate or remote, is then for us the whole of our conception of the object, so far as that conception has positive significance at all. (1907a/1991, pp. 23–24)

Mach's maxim for scientific theorizing agrees with Peirce and James that only differences that make a discernible difference matter for only they can make a difference to how we deal with the challenges to our survival or, less dramatically, our welfare.[42]

---

[39] Fittingly, Mach also gave an evolutionary-economical rendition of Hume's critique of the idea of causation as necessary connection in (1883/1960, p. 581) and a related dissolution of the nominalism-realism dispute in (1896/1986, p. 383).

[40] Mach fully recognized that the principle of economy in science possessed a socio-historical dimension: "The first real beginnings of science appear in society, particularly in the manual arts, where the necessity for the communication of experience arises" (Mach 1882/1943, p. 191). How far this anticipates the so-called Zilsel thesis cannot be considered here.

[41] See Holton (1992/1993, p. 11).

[42] To be sure, there are differences to be noted between Peirce's logical, James's psychological and Mach's methodological approach to the principles they formulated, but their convergence is undeniable and deserves to be noticed.

As Mach put it in *Principles of the Theory of Heat* a copy of the second edition of which he sent to James in 1900 and which James read two years later[43]:

> The character and course of development of science becomes more intelligible if we keep in mind the fact that science has sprung from the needs of practical life, from provision for the future, from techniques. ... The investigator strives for the removal of intellectual discomfort; he seeks a *releasing thought*. The technician wishes to overcome a practical discomfort; he seeks a *releasing construction*. Any other distinction between discovery and invention can scarcely be made. (1896/1986, p. 407)

The extent to which Mach's "positivism" was indeed a "pragmatism" as far as science was concerned can hardly be rendered plainer. As far as scientific method was concerned, James's pragmatism had little to add here.

6 Further Instances

It would be interesting to consider in this connection other German and Austrian philosopher-scientists and thinkers associated with the *Gesellschaft für positivistische Philosophie*, to see how widely my thesis holds.[44] James himself was fond of citing as inspiration the chemist and leader of the German Monist League Wilhelm Ostwald. Here I can only sketch the cases of one further scientist and one philosopher of note.

Support for a broadly naturalist-pragmatist approach to the understanding of science can also be found amongst representatives of the other side of the dispute about physical atomism that separated Mach and energeticists like Ostwald from realists like Max Planck. To be sure, not Planck himself (for whom recognition of the evolutionary origins of science did nothing to undermine its need for realistic metaphysics),[45] but Ludwig Boltzmann (who agreed with Mach's disdain of metaphysics, calling it a "spiritual migraine"). For instance, Boltzmann wrote:

> What leads to correct deeds is true. That is why I do not regard technological achievements as unimportant by-products of natural science but as logical proofs. Had we not attained these practical achievements, we should not know how to infer. Only those inferences are correct that lead to practical success. (1905a/1974, 192)

[43] See the letter from James to Mach of 17 June 1902 in Thiele (1978, p. 171).

[44] For a translation of the "Appeal" for the formation of this society, launched by Mach's acolyte Joseph Petzoldt "between late 1911 and summer 1912" and signed amongst others by Mach, Jerusalem, Schiller, and positivist philosophers like Schuppe and Ziehen as well as by luminaries like Hilbert, Einstein and Freud, see Holton (1992/1993, pp. 12–15).

[45] See Planck (1910).

Very much like Mach, Jerusalem and pragmatists like Schiller, Boltzmann was prepared to go all the way and did not shy away from psychologism.[46] Moreover, his evolutionism was of a decidedly pragmatic nature. Having died prior to the publication of James's *Pragmatism*, however, Boltzmann did not leave an assessment of this philosophy and his account of his travels in the United States in "Reise eines deutschen Professors ins Eldorado" (1905b) does not mention pragmatism either. His pragmatic evolutionism represents a convergence with pragmatism *avant la lettre* (albeit only partial as in the case of Mach).

Hans Vaihinger, an eminent Kant-scholar and founder of the Kant Society as well as the *Kant-Studien*, represents the opposite type of case in many respects. Vaihinger is generally reported as having written his widely read *Philosophy of the As If*, published in 1911, already in 1876–1877. What precisely dates back as far as that, however, is only Part 1, "Foundational Principles" (with some stylistic corrections); Part 2, "Special Investigations", and Part 3 on "Historical Anticipations", were written between 1906 and 1911 (Part 2 on the partial basis of earlier notes).[47] The published text therefore represents a considered reworking of ideas going back to the late 1870 s, but the central idea of his so-called fictionalism remained the same.

Vaihinger's self-styled "positivistic idealism" or "idealistic positivism" (he considered both terms interchangeable) held that not only do our values and ideals involve "fictions" whose ultimate function is to serve life, but also that the cognitive categories with which humans comprehend the world do not portray real existing types of entities – yet that despite their fictional nature they are indispensible and valuable. What Vaihinger stressed over and over was that the function of thought of orienting us in reality must not be mistaken for revealing the nature of reality to us. "One has to remember that representations in their entirety do not have the task of delivering a copy of reality – this is altogether impossible – but that they are an instrument to ease our orientation in it" (1911, p. 22, trans. TU). Vaihinger stressed that this holds not only for directly practical thought, but also for science.

That Vaihinger's work, which originally integrated the influence of Darwin, of Steinthal's and Wundt's psychologies and of Lange's critique of materialism, bore a certain affinity with pragmatism was widely noted when it was first published.[48] It is significant therefore that one of the "four moments" in the changed intellectual atmosphere which prompted Vaihinger to publish his early work some thirty years later (besides the "voluntarism" associated with Paulsen and Wundt, the "biological theory of knowledge" associated with Mach and Avenarius, and Nietzsche's philosophy) was "the pragmatism recently come to prominence" – within which Vaihinger then distinguished a "critical" from an "uncritical" version defending the former's opposition to "one-sided intellectualism and rationalism"

[46] "What then will be the position of the so-called laws of thought in logic? Well, in the light of Darwin's theory, they will be nothing else but inherited habits of thought." (Boltzmann 1905a/1976, 194)

[47] See Vaihinger (1921, pp. 192 and 194–195).

[48] See, e.g., Jerusalem (1912) and Jacoby (1912c).

against the latter's shallow utilitarianism (1911, p. ii–iv, trans. TU). Here again pragmatism itself was entirely innocent of influence on thinking with which it later was perceived to partially converge.

To summarise the thesis here developed for a fuller assessment on another occasion. A homegrown pragmatism of sorts had arisen (though clearly not under this name) among a few philosophers and scientists in Germany and Austria who were deeply impressed by Darwin's evolutionary theory and prepared to understand their own investigations within that framework to the extent that they embraced psychologism (or worse).[49] To the limited extent that there was a positive reception of American pragmatism by Central European philosophers and scientists prior to World War I, it was one that sprang from the recognition on their part that pragmatism agreed with conclusions they themselves had arrived at independently.

References

Baumgarten, E. 1938. *Die geistigen Grundlagen des amerikanischen Gemeinwesens. Bd. II. Der Pragmatismus: R.W. Emerson, W. James, J. Dewey*. Frankfurt a.M.: Klostermann.

Blackmore, J., and K. Hentschel (eds.). 1985. *Ernst Mach als Aussensseiter. Machs Briefwechsel über Philosophie und Relativitätsphilosophie mit Persönlichkeiten seiner Zeit*. Vienna: Braumüller.

Boltzmann, L. 1905a. Über eine These Schopenhauers. In *Populäre Schriften*, 385–402. Leipzig: Barth. Trans. 1974. On a thesis of Schopenhauer. In L. Boltzmann, *Theoretical physics and philosophical problems*, ed. B. McGuinness, 85–198. Dordrecht: Reidel.

Boltzmann, L. 1905b. Reise eines deutschen Professors ins Eldorado. In *Populäre Schriften*, 403–435. Leipzig: Barth.

Capek, K. 1918. *Pragmatismus.cili Filosofie praktickeho zivota*. Prague: F. Topic, 2nd rev. ed. 1925.

Dahms, H.-J. 1987. Aufstieg und Ende der Lebensphilosophie: Das philosophische Seminar der Universität Göttingen zwischen 1917 und 1940. In *Die Universität Göttingen unter dem Nationalsozialismus*, ed. H. Becker, H.-J. Dahms, and C. Wegeler. Munich: Saur. In 2nd enlarged ed. 1998, 287–317.

Dahms, H.-J. 1992. Positivismus und pragmatismus. In *Science and subjectivity*, ed. D. Bell and H. Vossenkuhl, 239–257. Berlin: Akademieverlag.

Dewey, J. 1930. *Demokratie und Erziehung.* Trans. E. Hylla. Breslau: Hirt. 2nd ed. Brauschweig: Westermann, 1949.

Dewey, J. 1931. *Die menschliche Natur, ihr Wesen und Verhalten.* Trans. P. Sakmann. Stuttgart/Berlin: Deutsche Verlagsanstalt.

Dewey, J., and W.H. Kilpatrick. 1935. *Der Projekt-Plan. Grundlegung und Praxis.* Trans. G. Schulz, E. Wiesenthal. Weimar: Böhlau.

Eckstein, W. 1935. *Wilhelm Jerusalem. Sein Leben und Wirken*. Vienna: Verlag von Carl Gerolds Sohn.

Elsenhans, T. (ed.). 1909. *Bericht über den III. Internationalen Kongress für Philosophie zu Heidelberg, 1. bis 5. September 1908*. Heidelberg: Carl Winter.

[49] Vaihinger's fictionalism embraced mathematics: "The basic concepts of mathematics ... are contradictory fictions. Mathematics rest on a wholly fictitious basis, even on contradictions." (1911, p. 71). This may not count as psychologism but is hardly an improvement on it.

Ferrari, M. 2005. Da sponda a sponda. 'Spirito tedesco' e 'tecnica americana'. In *Politiche della Tecnica. Immagini, Ideologie, Narrazioni*, ed. M. Neda, 189–212. Genova: Name.
Ferrari, M. 2010. Well, and pragmatism? Comment on Michael Heidelberger's paper. In *The present situation in the philosophy of science*, ed. F. Stadler et al., 75–83, Dordrecht: Springer.
Ferrari, M. 2014. Pragmatism and European philosophy: The French-Italian connection, Chapter 43. In *New directions in the philosophy of science*, ed. M.C. Galavotti, D. Dieks, W.J. Gonzalez, S. Hartmann, Th. Uebel, and M. Weber. Dordrecht: Springer.
Flournoy, T. 1911. *La Philosophie de William James*. Saint Blaise: Foyer Solidariste.
Frank, P. 1929–1930. Was bedeuten die gegenwärtigen physikalischen Theorien für die allgemeine Erkenntnislehre? *Die Naturwisswenschaften* 17(1929): 971–977 and 987–994, also *Erkenntnis* 1(1930): 126–157. Trans. 1949. In Frank, P. *Modern science and its philosophy*, 90–121. Cambridge, MA: Harvard University Press.
Gehlen, A. 1941. *Der Mensch. Seine Natur und seine Stellung in der Welt*. Berlin: Junker und Dünnhaupt. Rev. ed. Bonn: Athenäum, 1950. Reprinted in 1958.
Haller, R. 1986. Emanuel Herrmann. Zu einem beinahe vergessenen Kapitel österreichischer Geistesgeschichte. In *Fragen zu Wittgenstein und Aufsätze zur österreuichischen Philoso*phie, 55–68. Amsterdam: Rodopi.
Holton, G. 1992. Ernst Mach and the fortunes of positivism. *Isis* 83: 27–69. Reprinted in Holton, G. 1993. *Science and anti-science*, 1–55. Cambridge, MA: Harvard University Press.
Jacoby, G. 1909. *Der Pragmatismus. Neue Bahnen in der Wissenschaftslehre des Auslands*. Leipzig: Dürr. Repr. (edited by Lars Mecklenburg) In *Deutsche Zeitschrift für Philosophie* 50 (2002): 603–629.
Jacoby, G. 1912a. William James und das deutsche Geistesleben. *Die Grenzboten* LXXI(5): 217.
Jacoby, G. 1912b. William James' Angriff auf das deutsche Geistesleben. *Die Grenzboten* LXXI(3): 109–115.
Jacoby, G. 1912c. Der amerikanische Pragmatismus und die Philosophie des Als Ob. *Zeitschrift für Philosophie und philosophische Kritik* 147: 172–184.
James, W. 1899. *Der Wille zum Glauben und andere populärphilosophische Essays.* Trans. Th. Lorenz. Stuttgart: Fromann.
James, W. 1900. *Psychologie und Erziehung. Ansprachen an Lehrer.* Trans. F. Kiesow. Leipzig: Engelmann, 2nd ed. 1908.
James, W. 1907a. *Pragmatism. A new name for some old ways of thinking*. London: Longmans, Green & Co. Reprinted in Buffalo: Prometheus, 1991.
James, W. 1907b. *Die religiöse Erfahrung und ihre Mannigfaltigkeit*. Trans. G. Wobbenner. Leipzig: Hinrichs, 2nd ed. 1914.
James, W. 1908. *Pragmatismus. Ein neuer Name für alte Denkmethoden.* Trans. W. Jerusalem. Leipzig: Klinkhardt. Repr. Hamburg: Meiner, 1977.
James, W. 1909. *Psychologie.* Trans. M. Dürr. Leipzig: Quelle & Meyer.
James, W. 1914. *Das pluralistische Universum.* Trans. J. Goldstein. Leipzig: Kröner.
Jerusalem, W. 1905. *Der kritische Idealismus und die reine Logik. Ein Ruf im Streite*. Vienna: Braumüller.
Jerusalem, W. 1908a. Der Pragmatismus. *Deutsche Literaturzeitung*, 25th January. Reprinted in Jerusalem 1925, 130–139.
Jerusalem, W. 1908b. Philosophenkongress in Heidelberg. *Die Zukunft*, 10th October, 55–61.
Jerusalem, W. 1908c. Vorwort des Übersetzers. In *Pragmatismus. Ein neuer Name für alte Denkmethoden*, ed. W. James, iii–x. Trans. W. Jerusalem. Leipzig: Klinkhardt. Reprinted in Hamburg: Meiner, 1977.
Jerusalem, W. 1909. Apriorismus und Evolutionismus. In *Bericht über den III. Internationalen Kongress für Philosophie zu Heidelberg, 1. bis 5. September 1908*, ed. T. Elsenhans, 806–815. Heidelberg: Carl Winter.
Jerusalem, W. 1910a. William James. *Die Zukunft*, 5th November. Reprinted in Jerusalem, W. 1925. *Gedanken und Denker. Gesammelte Aufsätze. Neue* Folge, 2nd ed., 154–159. Vienna: Braumüller.

Jerusalem, W. 1910b. Preface to the English translation. In *Introduction to philosophy*, ed. W. Jerusalem, v–vii. New York: Macmillan.
Jerusalem, W. 1912. Die Logik des Unlogischen. *Die Zukunft*, 25th May. Reprinted in Jerusalem W. 1925. *Gedanken und Denker. Gesammelte Aufsätze. Neue* Folge, 2nd ed., 173–186. Vienna: Braumüller.
Jerusalem, W. 1913. Zur Weiterentwicklung des Pragmatismus. *Deutsche Literaturzeitung* 34: cols. 3205–3226.
Jerusalem, W. 1925. *Gedanken und Denker. Gesammelte Aufsätze. Neue Folge*, 2nd ed. Vienna: Braumüller.
Kleinpeter, H. 1911/1912. Der Pragmatismus im Lichte der Machschen Erkenntnislehre. *Wissenschaftliche Rundschau* 2(20): 405–407.
Kleinpeter, H. 1913. *Der Phänomenalismus. Eine naturwissenschaftliche Weltanschauung*. Leipzig: Barth.
Kusch, M. 1995. *Psychologism. A case study in the sociology of scientific knowledge*. London: Routledge.
Mach, E. 1872. *Die Geschichte und die Wurzel des Satzes von der Erhaltung der Arbeit*. Prague. Trans. *History and root of the principle of the conservation of energy*. Chicago: Open Court, 1911.
Mach, E. 1882. *Die ökonomische Natur der physikalischen Forschung*. Leipzig. Trans. The economical nature of physical inquiry. In E. Mach, *Popular scientific lectures*. Chicago: Open Court, 1893, 5th ed. 1943, 186–213.
Mach, E. 1883. *Die Mechanik in ihrer Entwicklung historisch-kritisch dargestellt*. Leipzig: Brockhaus. 9th ed. 1933. Trans. *The science of mechanics*. Chicago: Open Court, 6th ed. 1960.
Mach, E. 1884. *Über Umbildung und Anpassung im natutrwissenschaftlichen Denken*. Vienna. Trans. On transformation and adaptation in scientific thought. In *Popular scientific lectures*. Chicago: Open Court, 1893, 5th ed. 1943, 214–235.
Mach, E. 1896. *Die Principien der Wärmelehre.* Leipzig: Barth. 2nd ed. 1900. Trans. *Principles of the theory of heat*. Dordrecht: Reidel, 1986.
Mach, E. 1905. *Erkenntnis und Irrtum*. Leipzig: Barth. Trans. *Knowledge and error*. Dordrecht: Reidel, 1976.
Metz, R. 1933. C.S. Peirce, collected papers vols I and II. *Kant-Studien* 38: 188–189.
Müller, G. 1931. Charles Peirce (1840–1914). *Archiv für Geschichte der Philosophie* 40: 227–238.
Oehler, K. 1977. Einleitung. In James, W. *Der Pragmatismus. Eineuer Name für alte Denkmethoden*, ed. K. Oehler, ix*–xxxv*. Hamburg: Meiner.
Peirce, C.S. 1878. How to make our ideas clear. *Popular Science Monthly* 12: 286–303. Orig. "Comment render nos idees claires", *Revue Philosophique* 7 (1879): 39–57. Reprinted in C.S. Peirce, *The Essential Peirce Vol. I*, ed. Peirce edition project. Bloomington: Indiana University Press, 1992, 124–141.
Perry, R.B. 1936. *The thought and character of William James. As revealed in unpublished correspondence and notes, together with his published writings. Volume II philosophy and psychology*. Boston: Little, Brown and Company.
Planck, M. 1910. Zur Machschen Theorie der physikalischen Erkenntnis. *Physikalische Zeitschrift* 11: 1180–1198.
Schiller, F.C.S. 1903. *Humanism. Philosophical essays*. London: Macmillan & Co.
Schiller, F.C.S. 1906. Jerusalem, Der kritische Idealismus und die reine Logik. *International Journal of Ethics* 16: 391–393.
Schiller, F.C.S. 1907. *Studies in humanism*. London: Macmillan & Co.
Schiller, F.C.S. 1911. *Humanismus. Beiträge zu einer pragmatischen Philosophie.* Trans. R. Eisler. Leipzig: Klinkhart (later Kröner).
Schnädelbach, H. 1984. *Philosophy in Germany 1831–1933*. Cambridge: Cambridge University Press.
Schneider, H.W. 1938. E. Baumgarten, Der Pragmatismus: R.W. Emerson, W. James, J. Dewey. *Journal of Philosophy* 35: 695–698 [signed H.W.S.].

Scholz, H. 1934. Collected papers von Ch. S. Peirce, Bd. I–II. *Deutsche Literaturzeitung* 55: cols. 392–395.

Scholz, H. 1936. Collected papers von Ch. S. Peirce, Bd. III–V. *Deutsche Literaturzeitung* 57: cols. 137–144.

Shook, J. 2004. F.C.S. Schiller and European pragmatism. In *A companion to pragmatism*, ed. J. Shook and J. Margolis, 44–53. Oxford: Blackwell.

Simmel, G. 1895. Ueber ein Beziehung der Selektionslehre zur Erkenntnistheorie. *Archiv für systematische Philosophie* 1: 34–45. Trans. 1982. On a relationship between the theory of selection and epistemology. In *Learning, development and culture*, ed. H.C. Plotkin, 63–72. New York: Wiley.

Simmel, G. 1900. *Die Philosophie des Geldes*. Leipzig: Duncker & Humblot. Repr. Frankfurt a.M.: Suhrkamp, 1989.

Stein, L. 1908. Der Pragmatismus. Ein neuer Name für alte Denkmethoden. *Archiv für Philosophie. II. Abteilung: Archiv für systematische Philosophie* 14: 1–39 and 145–188.

Thiele, J. (ed.). 1978. *Wissenschaftliche Kommunikation. Die Korrespondenz Ernst Machs*. Kastellaun: Henn.

Uebel, T. 2012. But is it sociology of knowledge? Wilhelm Jerusalem's 'Sociology of Cognition' in context. *Studies in East European Thought* 64: 265–299.

Uebel, T. 2014. American pragmatism and the Vienna circle: The early years. *Journal for the History of Analytic Philosophy* 3(3).

Vaihinger, H. 1911. *Die Philosophie des Als Ob*. Leipzig: Meiner.

Vaihinger, H. 1921. Wie die Philosophie des Als Ob entstand. In *Die deutsche Philosophie der Gegenwart in Selbstdarstellungen. Zweiter Band*, ed. R. Schmidt, 175–203. Leipzig: Meiner.

Vorbrodt, G. 1913. W. James' Philosophie. *Zeitschrift für Philosophie und Philosophische Kritik* 151: 1–27.

Vorovka, K. 1929. *Americka Filosofie*. Prague: Bohumil Janda.

Walton, C., and J.P. Anton. 1974. Biographical sketch of Herbert W. Schneider. In *Philosophy and the civilizing arts essays presented to Herbert W. Schneider*, ed. C. Walton and J.P. Anton, xi–xxii. Athens: Ohio University Press.

New Prospects for Pragmatism: Ramsey's Constructivism

Maria Carla Galavotti

1 Ramsey's Pragmatism

Ramsey is often deemed a pragmatist, mainly due to his conception of truth, theories and scientific laws. In his article "Pragmatism" for the *Dictionary of the History of Ideas*, Philip Wiener includes Ramsey among those who support "a *relativistic* or contextualistic conception of reality and values in which traditional eternal ideas of space, time, causation, axiomatic truth, intrinsic and eternal values are all viewed as relative to varying psychological, social, historical, or logical contexts" (Wiener 1973, p. 553). In a similar vein, in *Meaning and Action. A Critical History of Pragmatism*, Horace Standish Thayer regards Ramsey as a pragmatist in connection with his conception of laws and theories, and calls attention to his "pragmatic criterion of the meaning of a sentence" (Thayer 1968, p. 311). Ramsey himself regarded this as the core of his pragmatist outlook, as testified by the following claim, contained in "Facts and Propositions": "the essence of pragmatism I take to be this, the meaning of a sentence is to be defined by reference to the actions to which asserting it would lead, or, more vaguely still, by its possible causes and effects" (Ramsey 1990a, p. 51). The centrality of action is indeed a crucial trait of pragmatism, as reflected by C.S. Peirce's well known maxim underpinning his theory of meaning, according to which "the only means of determining and clarifying the sense of an assertion consists in indicating what particular sort of experience one thereby intends to affirm will be produced, or would be produced, in given certain circumstances" (Peirce, "How to Make our Ideas Clear", in (1934), 5.402). A number of references to Peirce and James in Ramsey's writings are evidence of how deeply he was influenced by pragmatism.

M.C. Galavotti (✉)
Department of Philosophy and Communication, University of Bologna,
Via Zamboni 38, 40126, Bologna, Italy
e-mail: mariacarla.galavotti@unibo.it

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_45,

Given that human action is guided by belief, belief is the fundamental constituent of Ramsey's pragmatism, which revolves around a view of man as an agent acting in the world. This is the gist of Ramsey's viewpoint, as it prompts his conception of probability, causality, time, scientific laws and theories, and above all his theory of knowledge. Such a "perspectival" approach has been a source of inspiration for the "agency" theory of causality developed by Huw Price, partly in collaboration with Peter Menzies.[1] In "Opinions and Chances", appearing in *Prospects for Pragmatism*, Simon Blackburn labels Ramsey's view "projectivism", summarized as follows: "we regard the world as richer or fuller through possessing properties and things which are in fact mere projections of the mind's own reactions: there is no reason for the world to contain a fact corresponding to any given projection" (Blackburn 1980, p. 175). The same article contains a discussion of Ramsey's version of subjectivism, which Blackburn contrasts with the "irresponsible and therefore inefficient brand" with which he "is wrongly associated" (*ibid*., p. 195). As Blackburn emphasized, Ramsey's perspective has nothing to do with the "anything goes" picture of subjectivism that is still quite widespread, but, let me add, quite wrong.

The next sections will focus on aspects of Ramsey's perspectivalism that have not received the attention they deserve. Attention will be called to the importance Ramsey ascribed to the notion of predictive success, and to his conviction that the notions of objective chance and probability in physics can, and should, be accounted for within the subjectivist outlook. Furthermore, it will be argued that Ramsey can be considered a pioneer of the probabilistic view of knowledge and epistemology that later became predominant.

2 Ramsey's Subjectivism

Together with Bruno de Finetti, Ramsey is unanimously considered the father of modern subjectivism, namely the view according to which probability, taken as the degree of belief actually entertained by someone in a state of uncertainty regarding the occurrence of an event, is a primitive notion, having a psychological foundation. So conceived, subjective probability requires an operative definition specifying how it can be measured. This can be achieved in a number of ways, including the well-known method of bets, according to which one's degree of belief in the occurrence of an event can be expressed by means of the odds at which one would be ready to bet. Of this method Ramsey observes that it "suffers from being insufficiently general, and from being necessarily inexact" (Ramsey 1990a, p. 68) mainly because of the problem of the diminishing marginal utility of money. In view of this and other shortfalls, Ramsey opts for an alternative measure in terms of preferences, grounded on the conviction that "we act in the way we think most likely to realize

[1] See Price's chapter in this volume.

the object of our desires, so that a person's actions are completely determined by his desires and opinions" (*ibid*., p. 69).

The cornerstone of the subjective interpretation of probability is the notion of *coherence – consistency* in Ramsey's terminology. In terms of bets, coherence ensures that, if used as the basis of betting ratios, degrees of belief should not lead to a sure loss. In more general terms, this means that coherent degrees of belief satisfy the laws of probability. Fully aware of this, in "Truth and Probability" Ramsey states that the laws of probability can be shown to be "necessarily true of any consistent set of degrees of belief. Any definite set of degrees of belief which broke them would be inconsistent in the sense that it violated the laws of preference between options. [. . .] If anyone's mental condition violated these laws, his choice would depend on the precise form in which the options were offered him, which would be absurd. He could have a book made against him by a cunning better and would then stand to lose in any event. We find, therefore, that a precise account of the nature of partial belief reveals that the laws of probability are laws of consistency" (*ibid*., p. 78).[2]

In this approach, the laws of probability "do not depend for their meaning on any degree of belief in a proposition being uniquely determined as the rational one; they merely distinguish those sets of beliefs which obey them as consistent ones" (*ibid*.). This means that for subjectivists coherence alone guarantees the rationality of degrees of belief. Consequently, unlike the upholders of other interpretations of probability – including logicists like Rudolf Carnap and Harold Jeffreys – subjectivists do not regard probability evaluations as univocally determined by evidence, namely they admit that the same body of evidence is compatible with different evaluations of probability. For a subjectivist there are no "true" probability values, nor does it make sense to talk of "unknown" probabilities: probability is always known, being the expression of the opinion of those who evaluate it.

Bruno de Finetti's claim that "probability does not exist"[3] amounts precisely to the denial of a conception of probability as uniquely determined, inherent to facts. Much too often misunderstood and taken to mean that for subjectivists all coherent probability evaluations are on a par, de Finetti's claim is fully compatible with the awareness that probability assessments present an objectivity problem, or, as I.J. Good puts it, the problem of devising good probability appraisers.[4] Though committed to an extreme version of subjectivism, de Finetti took this problem very seriously and gave substantial contributions to its solution, which according to a widespread tendency is given in terms of *calibration*. In broad terms, the notion of calibration is essentially based on the idea that the goodness of probability appraisers depends on their predictive success. This idea, later developed by a

[2]The link between probability and degree of belief provided by coherence was discovered at about the same time by Ramsey and de Finetti, working independently. See Galavotti (1991) and Gillies (2000) for more on this.

[3]De Finetti wanted this claim printed in capital letters in the Preface of the English edition of his *Theory of Probability* (see de Finetti 1970/1975).

[4]See Good (1965) and Good et al. (1962).

number of statisticians including Leonard Jimmie Savage and many others,[5] is already entertained by Ramsey, who in comparing Keynes and Wittgenstein holds that: "a type of inference is reasonable or unreasonable according to the relative frequency with which it leads to truth and falsehood. Induction is reasonable because it produces predictions which are generally verified, not because of any logical relation between its premiss and conclusion" (Ramsey 1991, p. 301). This passage is evidence that Ramsey anticipated subsequent literature in this important respect.

The notion of success underpins the justification of induction Ramsey puts forward in "Truth and Probability". In his words: "induction is [. . .] a useful habit, and so to adopt it is reasonable. All that philosophy can do is to analyse it, determine the degree of its utility, and find on what characteristics of nature it depends. An indispensable means for investigating these problems is induction itself, without which we should be helpless" (Ramsey 1990a, p. 93). To which he adds: "This is a kind of pragmatism: we judge mental habits by whether they work, i.e. whether the opinions they lead to are for the most part true, or more often true than those which alternative habits would lead to" (*ibid.*).

In order to clarify subjectivists' attitude to probability evaluation, it is worth recalling de Finetti's distinction between the *definition* and the *evaluation* of probability, which he regards as different concepts, not to be conflated. Probability is *defined* as the *degree of belief* "as actually held by someone, on the ground of his whole knowledge, experience, information" (de Finetti 1968, p. 45) regarding an event whose outcome is uncertain. As this passage emphasizes, the evaluation of probability must take into account all available evidence, including frequencies and symmetries characterizing the events under consideration. However, it would be mistaken to put these elements, which are useful ingredients of the evaluation of probability, at the core of its definition. De Finetti believes that such a mistake affects the classical, logical and frequentist interpretations, whose proponents choose a single ingredient of probability evaluation to then place it at the core of its definition, associating probability with a unique function determined on the basis of that ingredient. By contrast, he regards the evaluation of probability as a complex procedure resulting from the concurrence of both objective and subjective elements: "every probability evaluation essentially depends on two components: (1) the objective component, consisting of the evidence of known data and facts; and (2) the subjective component, consisting of the opinion concerning unknown facts based on known evidence" (de Finetti 1974, p. 7). The objective component of probability judgments, namely factual evidence, is in many respects context-dependent: evidence must be collected carefully and skilfully, its exploitation depending on what elements are deemed relevant to the problem under consideration, and enter into the evaluation of probabilities. Therefore, the collection and exploitation of factual evidence involves subjective elements of various kinds. Equally subjective, according to de Finetti, is the decision on how to let belief be influenced by objective elements. Typically, one relies on information regarding frequencies. The

[5] See Savage (1971); on the notion of calibration see Dawid and Galavotti (2009).

interaction between degrees of belief and frequencies is reflected by the result that is often mentioned as de Finetti's "representation theorem", a combination of Bayes' rule and exchangeability. By showing that the adoption of Bayes' rule in conjunction with exchangeability leads to a convergence between degrees of belief and frequencies, de Finetti took a decisive step towards the edification of modern subjectivism, by indicating how subjective probability can be applied to statistical inference.

Although this crucial step was actually made by de Finetti, there is evidence that Ramsey knew the property of exchangeability, of which he must have heard from William Ernest Johnson's lectures.[6] Evidence for this claim is provided by his note "Rule of Succession", where use is made of the notion of exchangeability, named "equiprobability of all permutations".[7] What Ramsey apparently did not see, and was instead grasped by de Finetti, is the usefulness of applying exchangeability to the inductive procedure, modelled upon Bayes' rule. Remarkably, in another note called "Weight or the Value of Knowledge",[8] Ramsey was able to prove that collecting evidence pays in expectation, provided that acquiring the new information is free, and shows how much the weight increases. This shows he had a dynamic view at least of this important process. As pointed out by Nils-Eric Sahlin and Brian Skyrms, Ramsey's note on weight anticipates subsequent work by Savage, Good, and others.[9]

Bruno de Finetti, himself a pragmatist, regarded science as a continuation of everyday life and never paid much attention to the use made of probability in science, holding that subjective probability can do the whole job. Only the posthumous *Filosofia della probabilità* includes a few remarks that are relevant to the point. There de Finetti admits that probability distributions belonging to scientific theories – he refers specifically to statistical mechanics – can be taken as "more solid grounds for subjective opinions" (de Finetti 1995/2008, p. 63). This leads to conjecture that late in life de Finetti entertained the idea that probabilities encountered in science derive a peculiar "robustness" from scientific theories.[10] However, de Finetti's theory of probability does not contain a detailed account of a notion of probability specifically devised for application in science, and the same holds for the notion of chance.

By contrast, Ramsey took both of these notions seriously and made room for them within his subjectivist perspective. His way to account for objective probability within the framework of subjective probability represents an important contribution to subjectivism, which has not received the attention it deserves. Ramsey's view

[6] Johnson was the first to devise the probabilistic property of exchangeability; see Galavotti (2005) for more on this.

[7] See Ramsey (1991, pp. 279–281).

[8] See Ramsey (1990b), also included in (1991, pp. 285–287).

[9] See Nils-Eric Sahlin's "Preamble" to Ramsey (1990b) and Skyrms (1990, 2006). See in addition Savage (1954) and Good (1967).

[10] This is argued in some detail in Galavotti (2001, 2005).

of chance revolves around the idea that this notion requires some reference to empirical regularities. Ramsey makes clear that chance cannot be defined simply in terms of frequencies. As pointed out in "Reasonable Degrees of Belief": "we sometimes really assume a *theory* of the world with laws and chances and mean not the proportion of actual cases but what is chance on our theory" (Ramsey 1990a, p. 97). The same point is emphasized in the note "Chance", criticizing the frequency-based views of chance put forward by Norman Campbell. Incidentally, this highlights Ramsey's attitude to frequentism. Contrary to the widespread opinion that he was a dualist who admitted of two notions of probability, namely subjectivist and frequentist, Ramsey was critical of frequentism, which he deemed inadequate because "there is [. . .] no empirically established fact of the form 'In n consecutive throws the number of heads lies between $n/2 \pm \varepsilon$ (n)'. On the contrary we have good reason to believe that any such law would be broken if we took enough instances of it. Nor is there any fact established empirically about infinite series of throws; this formulation is only adopted to avoid contradiction by experience; and what no experience can contradict, none can confirm, let alone establish" (*ibid*., p. 104). To Campbell's frequentist account of chance, Ramsey opposed the view that "chances must be defined by degrees of belief" (*ibid*.). In his words: "chances are degrees of belief within a certain system of beliefs and degrees of belief; not those of any actual person, but in a simplified system to which those of actual people, especially the speaker, in part approximate" (*ibid*.). Ramsey stresses that chances "must not be confounded with frequencies", for the frequencies actually observed do not necessarily coincide with them: the chance of a roulette hitting the 0 is the same, irrespective of the fact that yesterday it never hit the 0 for the whole day. Unlike frequencies, chances can be said to be *objective* in two ways. First, to say that a system includes a chance value referred to a phenomenon, means that the system itself cannot be modified so as to include a pair of deterministic laws, ruling the occurrence and non-occurrence of the same phenomenon. Second, chances can be said to be objective "in that everyone agrees about them, as opposed e.g. to odds on horses" (*ibid*., p. 106).

Having defined chance, Ramsey goes on to deal with the notion of *probability in physics*, which he regards as chance referred to a more complex system, namely to a system making reference to scientific laws and theories. Physical probabilities can be regarded as *ultimate chances* in the sense that within the theoretical framework in which they occur there is no way of replacing them with deterministic laws. The assessment of physical probabilities is constrained by scientific theories, and their objective character descends from the objectivity ascribed to theories that are universally accepted. These are part of a strong system, supported by a good deal of evidence from experience. Ultimately, reference is made to the "true scientific system" which is "uniquely determined" because "long enough investigation will lead us all to it" (*ibid*., p. 161).[11]

[11] For more on Ramsey's notion of chance see Galavotti (1995).

Ramsey's view of chance and probability in physics is inextricably intertwined with his conception of theories, truth and knowledge in general. Notably, he accounts for the truth of theories in pragmatical terms holding the view, whose paternity he attributes to Peirce, that theories which gain "universal assent" in the long run are accepted by the scientific community and taken as true (*ibid.*). What he calls the "true scientific system" is accounted for in similar terms, namely as the system to which the opinion of everyone, grounded on experimental evidence, will eventually converge. As stated in "General Propositions and Causality", according to this pragmatically oriented view chance attributions – like all general propositions belonging to theories, including causal laws – are not to be taken as propositions, but rather as "variable hypotheticals", or "rules for judging" that "form a system with which the speaker meets the future" (*ibid.*, p. 149).

By grounding the notions of chance and physical probability on scientific laws and theories Ramsey attributes them a perspectival character, because laws and theories are themselves accounted for in a perspectival fashion. Therefore, perspectivalism is the common ground in which probability, chance, probability in physics and laws are all rooted. As argued in the next section, the same holds for Ramsey's concept of knowledge.

3 Knowledge

Ramsey's perspectival and constructivist view of knowledge departs considerably from the widespread conception of knowledge as justified true belief, discussed by Bertrand Russell (1912) in *The Problems of Philosophy*. Ramsey mentions Russell's work in the note "Knowledge" (1929), which opens as follows: "I have always said that a belief was knowledge if it was (i) true, (ii) certain, (iii) obtained by a reliable process" (Ramsey 1990a, p. 110). Ramsey then adds that the term "process" is unsatisfactory and for this reason condition (iii) should be dropped and substituted by the claim that belief qualifies as knowledge if it is "formed in a reliable way". He then holds that "we say 'I know' [. . .] whenever we are certain, without reflecting on reliability. But if we did reflect then we should remain certain if, and only if, we thought our way reliable. [. . .] For to think a way reliable is simply to formulate in a variable hypothetical the habit of following the way" (*ibid.*). This – only apparently slight – change in the traditional conditions imposed on belief for it to qualify as knowledge contains a remarkable novelty, because by putting the matter in these terms Ramsey grounds knowledge on the possibility of establishing a strong link between belief and success. As described by Nils-Eric Sahlin in an insightful analysis of Ramsey's position, for him "a belief is knowledge if it is obtained by a reliable process and if it always leads to success" (Sahlin 1990, p. 93).[12] Within

[12] See also Sahlin (1991).

Ramsey's perspective, knowledge is intertwined with the notion of causality, which is likewise grounded on reliable belief, namely on the kind of belief that is conducive to success.

Ramsey's last writings contain a few hints disclosing some criteria for the reliability of knowledge. First of all, in "General Propositions and Causality" he calls attention to experimentation, to which he ascribes a crucial role because it is by means of experimentation that beliefs are formed, and the weight of probabilities can be increased (Ramsey 1990a, p. 161). A further means for the formation of knowledge is provided by statistics, whose "significance is in suggesting a theory or set of chances" (see the note "Statistics" in 1990a, p. 102). Statistics is also the tool to perform causal analysis: "we find that chances are not what we expect, therefore the die is biassed" (*ibid*., p. 103).

A number of authors including Nils-Eric Sahlin, Jérôme Dokic and Pascal Engel[13] call attention to the fact that Ramsey's view of knowledge avoids altogether the well known difficulties that go under the name of the "Russell-Gettier problem", which affect the traditional view of knowledge. This is due to the fact that he grounds knowledge not only on the kind and amount of information backing our beliefs, but also on its reliability, which is itself grounded on past success.

The same authors, and others such as Simon Blackburn and Hugh Mellor, envisage a form of realism behind Ramsey's viewpoint. Blackburn speaks of "quasi-realism" in connection with Ramsey's projectivism, and in a similar vein Sahlin refers to Ramsey's conception of knowledge as obtained by a reliable process: "why do we regard some general beliefs as better habits qua basis for action than others? It is not because they are backed up by more evidence; that they have proved successful in the past. It is because we believe that there are underlying reliable processes or mechanisms accounting for our habits" (Sahlin 1991, p. 147). A discussion of Ramsey's attitude to realism falls beyond the scope of this paper.

The constructive attitude characterizing Ramsey's perspective is noteworthy. The criteria for reliability of knowledge he invokes, endowed with a distinctive pragmatist flavour, point to a view of epistemology and philosophy of science that regards probability as an essential ingredient of science – and human knowledge at large – and induction as the fundamental component of scientific method. By grounding knowledge on experimentation and statistics Ramsey anticipates the constructivist approach later developed by authors like Patrick Suppes, who takes "experimental methodology, including measurement, and statistical techniques for constructing and abstracting empirical structures in all domains of science" as the essential ingredient for the representation of scientific theories (Suppes 1988, p. 23).[14]

[13] See Dokic and Engel (2001).

[14] See also Suppes (1993, 2002).

4 Ramsey, Jeffreys and the Constructivist Approach

Ramsey's pragmatist and constructivist approach is shared by his contemporary the geophysicist and probabilist Harold Jeffreys, whose ideas may have influenced him.[15] Jeffreys embraces the logical notion of probability and holds that probability is uniquely constrained by evidence. At the same time, he shares some features of subjectivism like the conviction that there are no "unknown" probabilities, and the tenet that "no 'objective' definition of probability in terms of actual or possible observations, or possible properties of the world, is admissible" (Jeffreys 1939/1961, p. 11).

Like Ramsey, Jeffreys believes that the notion of probability arising in physics can be accounted for within the conceptual framework of epistemic probability. Critical of the frequentist interpretation of probability, Jeffreys holds that physical probability "can be discussed in the language of epistemological probability" (Jeffreys 1955, p. 284). In discussing quantum mechanics, taken as exemplifying a theory whose account of phenomena is irreducibly probabilistic, he calls attention to those branches of physics where "some scientific laws may contain an element of probability that is intrinsic to the system" (*ibid.*). Unlike the probability – or rather the chance – that a fair coin falls heads, intrinsic probabilities are grounded on theories: "intrinsic probabilities belong to the theory itself" (*ibid.*). The strict similarity between the position Jeffreys and Ramsey adopt in this connection should not pass unnoticed: for both of them physical probability derives its objective character from the theories accepted by the scientific community.

The two authors also take a similar stance regarding chance, which they both regard as the limiting case of everyday assignments. For Jeffreys chance occurs when "given certain parameters, the probability of an event is the same at every trial, no matter what may have happened at previous trials" (Jeffreys 1931/1957, p. 46). A remarkable aspect of Jeffreys' perspective is his conviction that the concepts of chance and scientific law acquire a definite meaning through scientific methodology. It is only "after the rules of induction have compared it with experience and attached a high probability to it as a result of that comparison" that a general proposition can become a law (Jeffreys 1939, p. 336). Moreover, in the procedure that leads to stating chances and laws lies "the only scientifically useful meaning of 'objectivity'" (*ibid.*). As put by Jeffreys: "I should query whether any meaning can be attached to 'objective' without a previous analysis of the process of finding out what is objective" (Jeffreys 1939/1961, p. 336). This process, stemming from experience, is inductive and requires the use of probability. When some hypothesis is supported by experience to such a degree that on its basis one can draw inferences whose probabilities are practically the same as if the hypothesis in question were certain, the association expressed by it can be asserted "as an approximate rule" (Jeffreys

[15] This conjecture is discussed in Galavotti (2003).

1937, p. 69). As one can see, Jeffreys' way of accounting for the notions of chance and law, as well as for the notion of objectivity, is strictly constructive.

A similar attitude characterises the author's view of causality, according to which causal analysis is based on statistical testing: it starts by considering all variation to be random and proceeds to detect correlations that allow for predictions and descriptions which are the more precise, the better their agreement with observations. This procedure leads to the assertion of laws, which are eventually accepted because "the agreement (with observations) is too good to be accidental" (*ibid.*). Within this perspective the principle of causality is "inverted": "instead of saying that every event has a cause, we recognize that observations vary and regard scientific method as a procedure for analysing the variation" (Jeffreys 1931/1973, p. 78). Also in this connection, there seems to be a striking analogy between the attitudes of Ramsey and Jeffreys, which can be described as forms of constructivism characterised by an awareness of the role played by statistical methodology in the formation of scientific knowledge.

As already observed, this attitude heralds the approach to epistemology and philosophy of science later developed by Suppes and many others. The attitude qualifies as bottom-up because it starts from specific problems arising within various disciplines and looks for possible solutions; epistemological notions are accounted for locally rather than forced into a general frame, and context becomes essential. This kind of approach has no room for the distinction between context of discovery and context of justification coined by logical empiricists. By contrast, a substantial continuity is established between the first steps of knowledge formation, including the methodology for the collection and representation of data, and the abstract models forming theories. Also in this connection, Ramsey and Jeffreys can be taken as forerunners of a turn that took place much later in the philosophy of science.[16]

References

Blackburn, S. 1980. Opinions and chances. In *Prospects for pragmatism*, ed. H. Mellor, 175–196. Cambridge: Cambridge University Press.

Cellucci, C. 2011. Classifying and justifying inference rules. In *Logic and knowledge*, ed. C. Cellucci, E. Grosholz, and E. Ippoliti, 93–106. Newcastle Upon Tyne: Cambridge Scholars Publishing.

Cellucci, C. 2013. Top-down and bottom-up philosophy of mathematics. *Foundations of Science* XVIII: 93–106. doi:10.1007/s10699-012-9287-6.

Dawid, A.P., and M.C. Galavotti. 2009. De Finetti's subjectivism, objective probability, and the empirical validation of probability. In *Bruno de Finetti, radical probabilist*, ed. M.C. Galavotti, 97–114. London: College Publications.

[16]Notably, in recent years constructivism has also attracted increasing attention among authors working on the foundations of mathematics and statistics. Recent contributions to constructivism in those areas have been put forward by the logician Carlo Cellucci – see his (2011, 2013) – and the statistician Christian Hennig – see his (2010). A discussion of their contribution is contained in Galavotti (2014).

de Finetti, B. 1968. Probability: The subjective approach. In *La philosophie contemporaine*, ed. R. Klibansky, 45–53. Firenze: La Nuova Italia.
de Finetti, B. 1970. *Teoria delle probabilità*. Torino: Einaudi. English edition: 1975. *Theory of probability*. New York: Wiley.
de Finetti, B. 1974. The value of studying subjective evaluations of probability. In *The concept of probability in psychological experiments*, ed. C.-A. Staël von Holstein, 1–14. Dordrecht/Boston: Reidel.
de Finetti, B. 1995. *Filosofia della probabilità*, ed. A. Mura. Milan: Il Saggiatore. English edition: 2008. *Philosophical lectures on probability*, ed. A. Mura. Dordrecht: Springer.
Dokic, J., and P. Engel. 2001. *Ramsey: Vérité et succès*. Paris: Presses Universitaires de France. English edition: 2002. *Frank Ramsey. Truth and success*. London: Routledge.
Galavotti, M.C. 1991. The notion of subjective probability in the work of Ramsey and de Finetti. *Theoria* LXII: 239–259.
Galavotti, M.C. 1995. F.P. Ramsey and the notion of 'chance'. In *The British tradition in the 20th century philosophy (proceedings of the 17th international Wittgenstein symposium)*, ed. J. Hintikka and K. Puhl, 330–340. Vienna: Hölder-Pichler-Tempsky.
Galavotti, M.C. 2001. Subjectivism, objectivism and objectivity in Bruno de Finetti's Bayesianism. In *Foundations of Bayesianism*, ed. D. Corfield and J. Williamson, 161–174. Dordrecht/Boston: Kluwer.
Galavotti, M.C. 2003. Harold Jeffreys' probabilistic epistemology: Between logicism and subjectivism. *The British Journal for the Philosophy of Science* LIV: 43–57.
Galavotti, M.C. 2005. *Philosophical introduction to probability*. Stanford: CSLI.
Galavotti, M.C. 2014. For a bottom-up approach to the philosophy of science. In *From a heuristic point of view. Essays in honor of Carlo Cellucci*, ed. E. Ippoliti and C. Cozzo, 199–215. Newcastle Upon Tyne: Cambridge Scholars Publishing.
Gillies, D. 2000. *Philosophical theories of probability*. London: Routledge.
Good, I.J. 1965. *The estimation of probabilities, an essay on modern Bayesian methods*. Cambridge, MA: The MIT Press.
Good, I.J. 1967. On the principle of total evidence. *The British Journal for the Philosophy of Science* XVIII: 319–321. Reprinted in I.J. Good 1983. *Good thinking. The foundations of probability and its applications*, 178–180. Minneapolis: University of Minnesota Press.
Good, I.J., A.J. Mayne, and J.M. Smith (eds.). 1962. *The scientist speculates. An anthology of partly-baked ideas*. New York: Basic Books.
Hennig, C. 2010. Mathematical models and reality: A constructivist approach. *Foundations of Science* 15: 29–48.
Jeffreys, H. 1931. *Scientific inference*. Cambridge: Cambridge University Press. Reprinted with "Addenda" 1937, 2nd modified ed. 1957, 1973.
Jeffreys, H. 1937. Scientific method, causality, and reality. *Proceedings of the Aristotelian Society, New Series* XXXVII: 61–70.
Jeffreys, H. 1939. *Theory of probability*. Oxford: Clarendon. 2nd modified ed. 1948, 1961, 1983.
Jeffreys, H. 1955. The present position in probability theory. *The British Journal for the Philosophy of Science* V: 275–289.
Peirce, C.S. 1934. *Collected papers of Charles Sanders Peirce*, ed. C. Hartshorne and P. Weiss. Cambridge, MA: Harvard University Press.
Ramsey, F.P. 1990a. *Philosophical papers*, ed. H. Mellor. Cambridge: Cambridge University Press.
Ramsey, F.P. 1990b. Weight or the value of knowledge. *The British Journal for the Philosophy of Science* XLI: 1–4.
Ramsey, F.P. 1991. *Notes on philosophy, probability and mathematics*, ed. M.C. Galavotti. Naples: Bibliopolis.
Russell, B. 1912. *The problems of philosophy*. London: William and Norgate.
Sahlin, N.-E. 1990. *The philosophy of F.P. Ramsey*. Cambridge: Cambridge University Press.
Sahlin, N.-E. 1991. Obtained by a reliable process and always leading to success. *Theoria* LVII: 132–149.
Savage, L.J. 1954. *Foundations of statistics*. New York: Wiley.

Savage, L.J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* LXVI: 783–801.

Skyrms, B. 1990. *The dynamics of rational deliberation*. Cambridge, MA: Harvard University Press.

Skyrms, B. 2006. Discovering 'weight, or the value of knowledge'. In *Cambridge and Vienna. Frank P. Ramsey and the Vienna Circle*, ed. M.C. Galavotti, 55–66. Dordrecht: Springer.

Suppes, P. 1988. Empirical structures. In *The role of experience in science*, ed. E. Scheibe, 23–33. Berlin/New York: Walter de Gruyter.

Suppes, P. 1993. *Models and methods in the philosophy of science: Selected essays*. Dordrecht/ Boston: Kluwer.

Suppes, P. 2002. *Representations and invariance of scientific structures*. Stanford: CSLI Publications.

Thayer, H.S. 1968. *Meaning and action. A critical history of pragmatism*. Indianapolis/New York: The Bobbs-Merrill Co.

Wiener, P. 1973. Pragmatism. In *Dictionary of the history of ideas*, vol. III, ed. P. Wiener, 551–570. New York: Charles Scribner's Sons.

Critical Realism in Perspective: Remarks on a Neglected Current in Neo-Kantian Epistemology

Matthias Neuber

1 What Is Critical Realism?

As a first approximation, critical realism can be characterized as an autonomous current in *transcendental revisionism*.[1] By "transcendental revisionism" I understand the attempt to reconcile the original Kantian doctrine with the developments of modern mathematics (the advent of non-Euclidean geometries in the first place) and modern physics (the advent of relativity theory in the first place). A closer look reveals that there were two dominant versions of transcendental revisionism in late nineteenth-/early twentieth-century philosophy in German-speaking countries. There was, firstly, the current of critical (or "logical") idealism of the so-called Marburg school of Neo-Kantianism. Defended by thinkers such as Hermann Cohen, Paul Natorp and Ernst Cassirer, the critical idealists' approach amounted primarily to a revision of the Kantian conception of the a priori. Thus, for example, Cassirer argued, both in his seminal *Substance and Function* (1910) and in his book on Einstein's theory of relativity (Cassirer 1921), for a replacement of the constitutive understanding of a priori principles by a purely regulative understanding. Accordingly, the original Kantian conception of static and absolutely valid a priori principles (like, for example, the principles of Euclidean geometry and Newtonian mechanics) became transformed into a more dynamical and relative conception of such principles. This transformation, in turn, allowed – at least on the critical idealist reading – to account for the revolutionary changes in mathematics and physics around 1900.[2]

[1] The following draws heavily on (Neuber 2011, 2012).

[2] For a detailed reconstruction of Cassirer's (and the Marburg school's) revisionism see, especially, (Friedman 2000, Ch. 7). See further (Neuber 2012, ch. 3).

M. Neuber (✉)
Department of Philosophy, University of Tübingen, Bursagasse 1, D-72070 Tübingen, Germany
e-mail: matthias.neuber@uni-tuebingen.de

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_46,

No doubt, the critical idealist movement was an important (and quite influential) attempt at revising Kant's theory of scientific knowledge. However, there was a second – less well-known – current of transcendental revisionism, namely the critical *realist* movement. As I will argue, this movement deserves more attention than it has received so far. Yet, for the moment it will suffice to make note of two issues: (1) The critical realists agreed with the critical idealists that the aim of philosophy consists in the critical reflection on the preconditions of scientific knowledge. This is the reason why both currents share the attribute "critical" and why both currents are to be attached to the more general project of a "scientific philosophy" (*wissenschaftliche Philosophie*).[3] (2) The critical realists disagreed with the critical idealists on the determination of the very *object* of science. While the critical idealists confined the realm of scientific (especially physical) objects to the progress of conceptual development – thereby implying that the object of scientific knowledge is not "given" (*gegeben*) but only "set as a task" (*aufgegeben*) – the critical realists insisted that the objects of science are *given as they are in themselves*. This sounds somewhat naïve, but we will see later that the issue was more complex and less direct. For the time being, however, it should be kept in mind that the assumption of knowable things-in-themselves was the peculiar feature of the critical realists' attempt at revising the Kantian conception of scientific knowledge.

2 Some Protagonists in the Debate

Taking a less well-known philosophical current in perspective means (among other things) to tell something about its outstanding proponents. In the case of critical realism this task is rather challenging because the movement as a whole was quite heterogeneous. The following overview should therefore be regarded as a first and (very) preliminary sketch in need of further elaboration.

To begin with, the one who might be seen as the "founder" of the critical realist movement was Alois Riehl (1844–1924). In his three-volume *Der philosophische Kriticismus und seine Bedeutung für die positive Wissenschaft* (1876, 1879, 1887), Riehl attempted to combine Kantian criticism with certain elements of British (especially Lockean) empiricism. As becomes obvious from the title, Riehl developed his ideas in direct confrontation with the "positive" (experiential) sciences. In this respect he was a typical representative of the idea of a scientific philosophy (see, in this connection, also Riehl 1883). Furthermore, it was Riehl who, in volume 1 of *Der philosophische Kriticismus*, defended the opinion that Kant's *Critique of Pure Reason* is intimately correlated with a "theory of reality" (*Wirklichkeitslehre*) (1876/1908, p. 562). Thus, in a section headed "Erscheinung, Ding an sich, Noumenon", Riehl argued for the reality of things-in-themselves, thereby refusing to interpret them as mere "limiting concepts" in the sense of Marburgian

[3]For the details of this project see (Richardson 1997, 2008); (Friedman 2001, ch. 1, 2012).

Neo-Kantianism (*id.*, p. 561).[4] For Riehl, things-in-themselves had to be taken ontologically seriously. More precisely, he saw in them the objective basis of spatio-temporal appearances.[5] Correspondingly, Riehl, in volume 3 of *Der philosophische Kriticismus*, argued that we acquire knowledge of things-in-themselves *via* our knowledge of appearances (1887/1926, pp. 164–165). We will come back to this point in the following section. In the present context, though, it is important to realize that Riehl characterizes the view that we acquire knowledge of things-in-themselves via knowledge of appearances explicitly as "critical realism" (*id.*, p. 163); furthermore, he distinguishes this "Copernican" attitude (as he puts it) from the "Ptolemaic" or "phenomenalistic" attitude (*ibid.*). Whereas according to the latter realism *precedes* criticism, the Copernican – critical realist – approach implies that realism *follows* criticism.[6]

A similar, though clearer and more explicit distinction can be found in the work of another proponent of the critical realist movement, Wilhelm Wundt (1832–1920). Wundt, one of the founding fathers of experimental psychology, published in 1896 an extended article titled "Über naiven und kritischen Realismus". In this paper, Wundt was primarily engaged with refuting the (as it was termed) philosophy of immanence (Schuppe, Schubert-Soldern, Rehmke), on the one hand, and "empiriocriticism" (Avenarius, Mach, Petzoldt), on the other. Since both schools of thought sought to base knowledge on the "immediately given", Wundt lumped them together under the label "naïve realism" (1896/1910, p. 265). From this, in his opinion wrong-headed, point of view he demarcated "critical realism" (*id.*), the characteristic feature of which he saw in the *emancipation* from the immediately given. In consequence, critical realism, in the view of Wundt, was committed not only to the reality of what was directly perceivable but also to the reality of atoms, electromagnetic fields and other inferred (theoretically postulated) entities. The essential tool of critically scrutinizing the positing of such entities was, according to Wundt, the systematic investigation of the *history* of science (*id.*, pp. 267–268). Given the liability of the historical method, the "Copernican attitude" in Wundt amounted to the "correction" of our immediate sensory experiences by the positing and critical scrutiny of theoretical, not directly perceivable constructs. That is, for Wundt, realism *follows* criticism insofar as it is dependent on a preceding analysis of the factual (objective) basis of our direct perceptual experience.[7]

---

[4]For their interpretation as mere "limiting concepts", see especially (Natorp 1910, ch. 1, §§ 4–6) and (Cassirer 1910, ch. 7).

[5]See (Riehl 1876/1908, p. 476): "It follows from Kant's theory, even if Kant did not expressly say so, that there must be a reason for every particular empirical determination of space and time in the object that appears."

[6]More precisely, naïve realism, according to Riehl, takes perceptual reality for granted without reflecting on its origins. It is, as might be said, the view of the "man in the street."

[7]It is noteworthy, in this connection, that Wundt's Belgian student Georges Dwelshauvers wrote some manner of a French short version of Wundt's original article. Dwelshauvers's contribution was titled "Réalisme naïf et réalisme critique" and also appeared in 1896.

The perhaps most articulated version of the critical realist program can be found in the work of Wundt's former student (and founder of the "Würzburg school of psychology") Oswald Külpe (1862–1915). Particularly interesting in this connection in his three-volume magnum opus *Die Realisierung: Ein Beitrag zur Grundlegung der Realwissenschaften* (1912, 1920, 1923). As the subtitle indicates, Külpe's aim is to deliver a "foundation" of the "real" (experiential) sciences. Thus, like Riehl and Wundt, Külpe collaborates in the project of realizing the idea of a scientific philosophy. Furthermore, his understanding of the term "critical" comes rather close to the one propounded especially by Wundt. In a more or less Lockean vein,[8] Külpe draws a principled distinction between subjective "sense qualities", on the one hand, and objective, non-perceptual "realities", on the other (1912, pp. 1–3). While, according to Külpe, "naïve" realism is confined to the analysis of the former, "critical" realism sets itself the task to critically scrutinize the theoretically postulated entities ("realities") of science (*id.*, p. 26). It is for this reason that critical realism, for Külpe, is identical with *scientific* realism (*wissenschaftlicher Realismus*) (*id.,* p. 45), which, in turn – as Külpe writes at another place – has the status of a "probable hypothesis" (Külpe 1910, p. 161).[9] Moreover, Külpe's attitude toward the original Kantian doctrine becomes sufficiently clear by his discussion of the following two questions: (1) "How is a positing [*Setzung*] of the real possible?" (1912, p. 4) (2) "How is a determination [*Bestimmung*] of the real possible?" (*id.*, p. 5) In Külpe's view, Kant only answered the first question. That is, he postulated the existence of things-in-themselves, but he did not say anything about their nature. We will see in the following section that Külpe himself is eager to tackle also the challenge of the second question by offering an elaborated theory of scientific *explanation*.[10]

To sum up so far, critical realism is best characterized as a late nineteenth/early twentieth-century variant of transcendental revisionism. It embraces two key aspects: (1) At the methodological level, critical realism forms part of the larger project of realizing the idea of a scientific philosophy. It is apt to speak of the *anti-metaphysical* (or *anti-speculative*) aspect in this connection.[11] (2) At the epistemological level, critical realism can be identified by the idea that Kantian things-in-themselves are knowable. Yet, it is this very idea that needs further clarification.[12]

---

[8]I have to thank Adam Caulton for helpful comments in this connection.

[9]It is interesting to note that the characterization of scientific realism as a "probable hypothesis" comes remarkably close to more current ("naturalistic") scientific realist positions. See, in this connection, e.g. (Boyd 1983), (Psillos 1999), (Sankey 2008).

[10]For further details of Külpe's general conception of "realization", see (Henckmann 1997).

[11]This anti-metaphysical impetus has no doubt to do with the mid-nineteenth century breakdown of speculative *Naturphilosophie*, especially in sense of Hegel and Schelling, and the parallel strengthening of the institutionalized (natural) sciences. For further details in this respect, see (Schnädelbach 1983, ch. 3) and (Neuber 2012, pp. 46–48).

[12]Again it should be stressed that the critical realist movement was, as indicated above, a rather complex movement. It is quite striking that among its defenders were many psychologists, such as (besides the already mentioned Wundt and Külpe) Gustav Störring (1860–1946), August Messer

3 The Knowability Thesis

The most straightforward definition of the *realist component* of critical realism was given by one of its leading advocates, the Munich philosopher and psychologist Erich Becher.[13] For Becher, "*[r]ealism is the doctrine that things-in-themselves are knowable*" (1914, p. 69). Now, from a "genuinely Kantian" perspective, this point of view is more than problematic. Without devolving into exegetical details, it can be stated that, for Kant, the assumption of (theoretically) knowable things-in-themselves was anathema. Thus, as early as in the "Transcendental Aesthetic" of his *Critique of Pure Reason*, Kant categorically declares:

> It is [. . .] indubitably certain and not merely possible or even probable that space and time [. . .] are merely subjective conditions of all our intuition, in relation to which therefore all objects are mere appearances and not things given for themselves in this way; about these appearances, further, much may be said *a priori* that concerns their form but nothing whatsoever about the things in themselves that may ground them. Kant ([1781/1787] 1998, B 66)[14]

Kant called this doctrine "transcendental *idealism*". As he puts it in the "Transcendental Dialectic" of his *Critique of Pure Reason*, transcendental idealism implies that "everything intuited in space or in time, hence all objects of an experience possible for us, are nothing but appearances, i.e., mere representations, which, as they are represented, as extended beings or series of alterations, have outside our thoughts no existence grounded in themselves." (B 518–519)

To be sure, Kant, at the same time, refuted (Berkeleyan) "dogmatic" and (Cartesian) "skeptical" (or "problematic") idealism (B 274–279 and B 377–380) and argued in favor of what he called – at least in the first edition of the *Critique of Pure Reason* – "empirical realism" (A 370). However, this sort of realism was explicitly restricted to knowledge of appearances and had therefore nothing to

(1867–1947), and Erich Becher (1882–1929). Other critical realists in the German-speaking area were, to mention just a few, Willy Freytag (1873–?), Max Frischeisen-Köhler (1878–1923), Bernhard Bavink (1879–1947), Victor Kraft (1880–1975), and Aloys Wenzl (1887–1967). Roughly at the same time there was a realist movement in the United States, including both the so-called "new" and the "critical" realists; for the details of this distinction, see Drake et al. (1920, p. vi) and (Sellars 1969, ch. 4). Among the latter were thinkers such as George Santayana (1863–1952), Arthur Lovejoy (1873–1952), and Roy Wood Sellars (1880–1973). The relation between the German and the American brand of critical realism has, it should be emphasized, still to be researched.

[13]Interestingly enough, it was originally Becher who was supposed to receive the prestigious Chair of the History and Philosophy of the Inductive Sciences at the University of Vienna. Since Becher, in September of 1921, declined the appointment; it was Moritz Schlick who, in August of 1922, eventually was awarded the chair. Furthermore, Becher and Schlick, since 1915, stood in correspondence, and it is obvious that (with the exception of the issue of vitalism) there was much agreement between their respective philosophical points of view. For further details in this connection, see (Engler and Neuber 2006, pp. 124–125) and (Friedl and Rutte 2008, pp. 11–14).

[14]All references beginning with either "A" or "B" are to the first or second edition of Kant's *Critique of Pure Reason* (1781–1787).

do with things-in-themselves. On Kant's account, no other form of realism was worth taking seriously. In particular what he called "*transcendental* realism", in his view, was doomed to fail. According to Kant, transcendental realism "makes mere representations into things in themselves" (B 519). The fatal consequence of this maneuver is that realism becomes indistinguishable from *empirical idealism*. Kant writes:

> [T]ranscendental realism necessarily falls into embarrassment, and finds itself required to give way to empirical idealism, because it regards the objects of outer sense as something different from the senses themselves and regards mere appearances as self-sufficient beings that are found external to us; for here, even with our best consciousness of our representation of these things, it is obviously far from certain that if the representation exists, then the object corresponding to it would also exist; but in our system [i.e. transcendental idealism; M.N.], on the contrary, these external things – namely, matter in all its forms and alterations – are nothing but mere representations, i.e., representations in us, of whose reality we are immediately conscious. (A 371–372)

The plausibility of Kant's conception notwithstanding, it is clear that the critical realist movement, according to Kant himself, would fall under the rubric of transcendental realism and should therefore be rejected. More precisely, Kant would argue that the critical/transcendental realist must problematically *infer* his putative object of experience, whereas a more empirical understanding tells us that – space and time being 'pure forms of intuition' – objects of experience can be *directly perceived*. Thus, in contrast to the critical/transcendental realist, "the transcendental idealist is an empirical realist, and grants to matter, as appearance, a reality which need not be inferred, but is immediately perceived (A 371)."[15] Or in other words: Transcendental idealism accounts for the needs of a non-metaphysical (non-inferential) and at the same time realist conception of empirical knowledge.[16]

However, there is no revision without (at least a certain degree of) demolition. That is, as soon as the attempt is made to modify certain aspects of the original Kantian doctrine, a more or less new and independent conception – partly or even entirely alien to the original – is the (intended) outcome. This holds true for Marburgian Neo-Kantianism, and it holds no less true for critical realism. Accordingly, transcendental revisionism in general should absolutely not be conflated with Kant exegesis.

---

[15]See, in this connection, also Kant's reflections on the "transcendental ideality" and "empirical reality" of space and time at (B 44 and B 52–53).

[16]It should be noted that Kant's argument for transcendental idealism is highly controversial. Having its roots in the late eighteenth- and nineteenth-century (Jacobi, Maimon, Trendelenburg, Fischer, etc.), the debate over transcendental idealism has found its continuation in the more recent contributions by (Strawson 1966), (Allison 1983), (Guyer 1987), and (Langton 1998). Whereas, for example, Allison explicitly aims at a "defense" of transcendental idealism, Guyer (following Strawson) critically argues that "Kant's argument for transcendental idealism from his theory of the forms of intuition does not express epistemological modesty but is rather the consequence of an exceedingly immodest interpretation of the necessity of synthetic *a priori* propositions" (1987, p. 369). As we will see later, Schlick had a rather similar objection against the transcendental idealist project.

As for the critical realist movement in particular, the assumption of knowable things-in-themselves was fairly diametrically opposed to Kant's original doctrine. As already indicated, the crux of the critical realist revision was the view that things-in-themselves can be known *via* our knowledge of appearances.[17] Riehl, for example, argued that (surprisingly, according to Kant) "mediate" (*mittelbare*) knowledge of things-in-themselves is possible (1876/1908, p. 479 and 1887/1926, p. 165). Similarly, Külpe held that appearances are ontologically "grounded" in things-in-themselves and that the latter can therefore be inferred from the former (1920, pp. 210–212). More precisely, Külpe was of the opinion that things-in-themselves *can be determined with respect to their very nature by abductive reasoning* (cf. *id.* and Neuber 2012, pp. 52–53). From the original Kantian perspective, this would be impossible, because the causality principle, according to Kant, does not apply to things-in-themselves, but only to intuited phenomena (cf. B xxvii–xxviii, B 232–256). However, within the critical realist conception, knowledge of things-in-themselves is possible since the Kantian faculty of intuition plays only the role of an "epistemic starting point" that we eventually abandon for the sake of purely conceptual knowledge. Thus, for example, Riehl writes in this connection:

> We see, in fact, how science reduces the content of experience to its lawlike elements, to what recurs in similar form, to what is accessible to quantitative determination and is thus expressible in numerical operations, in short, to the *conceivable*. Everything else is an object, not of *conceiving*, but of immediate *acquaintance*, and hence of feeling, sensation and perception. (1879/1925, p. 221)

Riehl's distinction between perceptual – i.e. everyday – "acquaintance" (*Wissen*) and quantitatively determined – i.e. scientific – "conceiving" (*Begreifen*) can hardly be overestimated in its historical significance; a very similar distinction can be found in the early work of Bertrand Russell.[18] More to the point, it was the early Schlick who, in his *General Theory of Knowledge*, explicitly took over Riehl's (and Russell's) distinction (cf. Schlick 1925/1974, p. 83) (we will return to this issue later). In the present context, it is important to realize that the critical realists thought that, by downgrading the role of intuition, the route to knowledge of things-in-themselves becomes free. It is for this reason that, for instance, Külpe straightforwardly *defined* knowledge as "conceptual coordination" (*begriffliche Zuordnung*) and declared: "For objective science, concepts are '*fixed coordinations*' between signs and signified objects." (1912, p. 226) By "signified objects", Külpe meant, not surprisingly, things-in-themselves.

It is interesting to note that the critical *idealists* also downgraded the role of the Kantian faculty of intuition. But they draw different consequences form

[17] As Rae Langton (1998, pp. 89–93) has convincingly pointed out, the view that knowledge of things-in-themselves can be acquired via knowledge of appearances (phenomena) has one of its roots in Kant's interpretation (and critique!) of Leibniz's version of idealism. See, in relation to this, esp. (B 60, B 323, B 326).

[18] See, in this connection, Russell's distinction between "knowledge by acquaintance" and "knowledge by description" in (Russell 1910–1911).

this downgrading. According to the critical idealists, it was not the realm of things-in-themselves that became accessible by viewing intuition only as an epistemic starting point (as I would term it). Rather, it was *the conceptual system of science itself* that essentially became the focus of analysis (see, e.g., Cohen 1902; Natorp 1910, esp. p. 95; Cassirer 1910; also Neuber 2012, pp. 135–136). In the background of this – plainly un-Kantian – shift of perspective stood, among other things, the development of modern ("unintuitive") mathematics (cf. Friedman 2005). In the case of the critical realists, however, *psychology* played a decisive role. Thus, for example, Külpe argued as early on as in his *Grundriss der Psychologie* of 1893 – and later other members of the Würzburg school of psychology (Ach, Bühler, Messer, etc.) would follow in his footsteps – for an experimentally supported theory of "intuition-free thinking" (*anschauungsfreies Denken*) that he, in the central work for his philosophical program (i.e. *Die Realisierung*), took as the basis for establishing a general ("semiotic") theory of knowledge of "realities" (which are essentially things-in-themselves).[19] Later, it will be made clear later that Schlick was heavily inspired by Külpe's theory.

One further aspect of the critical realist program deserves consideration. It has probably become clear by now that in the view of the critical realists, the objects of science had the status of things-in-themselves. *But what was the status of things-in-themselves?* The critical realist answer to this question is quite traditional. It amounts to the (rather Aristotelian) view that what is "really real" are *substances*. Thus, for example, Becher equated "the thing" with substance and characterized the latter as the "bearer of attributes" which has existence and identity independently of any possible attribute (1915, p. 10). In a similar vein, Riehl defined the thing-in-itself as "the persistent" (*das Beharrliche*) and qualified this definition by adding that "matter and force are the substantial of outer experience" (1887/1926, p. 66). Moreover, Riehl devoted a whole chapter of the second volume of *Der philosophische Kriticismus* to the connection of substance, matter and force (1879/1925, pp. 303–322), arguing, among other things, that physics and chemistry "deliver the proof of the quantitative persistence of matter" (*id.*, p. 313). The most extended critical realist discussion of the substantialist account of things-in-themselves can (as far as I can see) be found in volume 3 of Külpe's *Die Realisierung* (1923, pp. 244–309). There, Külpe points out that things-in-themselves, i.e. substances, can be determined as the independently existing, unchangeable and persistent basis of the scientific explanation of the behavior of observable phenomena. Being essentially inferred, substantial things-in-themselves, for Külpe, are entities like atoms, molecules and energies (*id.,* pp. 245–246).[20] It is on the basis of the postulation of these entities that appearances become explainable in terms of

[19](Lindenfeld 1978) gives a fairly instructive account of the relation between psychology and philosophy in Külpe's work.

[20]There is sufficient evidence that Kant would *not* have accepted the characterization of atoms, molecules, etc. as substantial things-in-themselves. See, in this connection, (Langton 1998, pp. 207–208).

scientific realism.[21] In fact, it is this sort of "inference to the best explanation" that should be seen as one of the most pivotal features of Külpe's general conception of "realization" (*id.*, pp. 308–309). On the whole, it can be said that Külpe, no less than Riehl (and Becher), "substantialized" the Kantian things-in-themselves.[22]

However, the critical realists' indebtedness to substantialist thinking might be regarded as somewhat reactionary. At any rate, it might be argued that at the time when the critical realists put forward their substantialist account of science and nature a viable alternative was already available. I think of Cassirer's *Substance and Function* here, and I dare claim that Cassirer indeed had a serious objection against all forms of substantialism. In Cassirer's view (as is widely known), the development of modern science moves from a (quasi-Aristotelian) substantialist understanding of concepts, such as "energy" or "atom", to a functional (or relational) understanding of such concepts (cf. Cassirer 1910, esp. Chap. 4). Without elaborating on the details of this theory, I confine myself here to indicating that Cassirer's point of view had tremendous impact on the subsequent history of thought within the German-speaking context.[23]

4 Schlick's Appropriation of the Critical Realist Program

Having inspected in some detail the critical realist program, we are now in a position to discuss what the early Schlick made out of it. In order to give some context to what follows, let us quote from Herbert Feigl, who – in an article titled "The Power of Positivistic Thinking" – wrote:

> [M]y original position, long before I decided to study at the University of Vienna, had been close to a critical realism which I had first formulated for myself in a rather unsophisticated manner when during my adolescence I was deeply impressed with the achievements of

[21] For a related, though more recent point of view, see Psillos's discussion of Richard Boyd's "explanationist defence of realism" in (Psillos 1999, pp. 78–81).

[22] A more detailed reconstruction of the critical realist program would have to take into consideration the notorious rivalry between the "two-aspects view" and the "two-worlds-view" of transcendental idealism. Unfortunately, for reasons of brevity, I am only able to allude here to the fact that Riehl tended toward the first interpretation, while Külpe tended toward the second. This, in a nutshell, meant that, for Külpe, things-in-themselves constituted an autonomous class of entities being causally responsible for the "affection" of observable appearances, whereas, for Riehl, things-in-themselves did *not* constitute a special class of entities but were merely the product of viewing appearances "in abstraction" from the conditions of sensibility. Riehl explicitly states that "the Critique teaches us to take the *same* object in two meanings, as appearance and as thing-in-itself" (1876/1908, p. 562). Other advocates of the two-aspects view are (Paton 1936), (Prauss 1974), and (Allison 1983). The two-worlds view, on the other hand, is defended by (Vaihinger 1881/1892), (Strawson 1966), (Guyer 1987). See further (Willaschek 2001), according to whom this whole discussion is essentially misguided.

[23] For an extended discussion of Cassirer's relationalist point of view, see (Neuber 2012, pp. 137–155). See further (Friedman 2005, pp. 73–77).

> astronomy, chemistry, and theoretical physics. I began reading the positivists, but also the Neo-Kantian Alois Riehl, the German critical realists Külpe, Becher, and Freytag during my last year in secondary school; and finally on the suggestion of a distant relative, a prominent professor of medicine (at the German University in Prague), I read two books by Schlick which had been highly recommended to him by no less a person than his old friend Albert Einstein. Schlick's *Allgemeine Erkenntnislehre* [. . .] struck me like a thunderbolt. In the beautiful lucid and magnificently penetrating book Schlick argued essentially for a critical realism, presenting trenchant objections to what he called the philosophies of immanence – that is, mainly the positions of Mach, Avenarius, and the early Russell. This, together with his views on the analytic nature of mathematical truth, his empiricist critique of Kant and the Neo-Kantians, and his profound understanding of modern science motivated me to become his student at the University of Vienna in 1922. But I was acutely distressed to witness Schlick's conversion to positivism in the late twenties. This conversion was largely due to the influence of Carnap and Wittgenstein. (1963/1981, pp. 38–39)

It must be recognized that Feigl's claim of a "conversion" is not uncontested. Ludovico Geymonat (1985), for example, has argued that many ideas that can be found in the work of the early Schlick reoccurred in his later – Viennese – writings as well as in the writings of Carnap and Wittgenstein, who were allegedly responsible for his "conversion". In the final section of this paper, it will be made clearer that there are good arguments in favor of Feigl's rather than Geymonat's assessment. Yet, for the time being, it is enough to emphasize that Schlick's early (pre-Viennese) point of view stood in the critical realist tradition.

That said, one should recognize that Schlick never *explicitly* characterized himself as a critical realist.[24] Accordingly, it appears plausible to characterize his early philosophical position as merely being inspired by the critical realist tradition rather than being critical realist *tout court*. On the interpretation I wish to offer, the early Schlick's position was somewhat of a hybrid of the critical realists' insistence on knowable things-in-themselves and the early Cassirer's *relationalism*.

Before going into the details of this interpretative proposal, a few words must be said concerning Schlick's appreciation of Kantian theoretical philosophy. In the first instance, Kant, according to Schlick, was one of the most important pioneers of the idea of a scientific philosophy. Thus, in his paper on "The Philosophical Significance of the Principle of Relativity", Schlick points out:

> We have known since the days of Kant that the only fruitful method of all theoretical philosophy consists in critical inquiry into the ultimate principles of the special sciences. Every change in these ultimate axioms, every emergence of a new fundamental principle, must therefore set philosophical activity in motion [. . .]. [T]he Kantian Critical Philosophy may itself be regarded as a product of the Newtonian doctrine of nature. It is primarily, or even exclusively, the principles of the exact sciences that are of major philosophical importance, for the simple reason that in these disciplines alone we do find foundations so firm and sharply defined, that a change in them produces a notable upheaval, which can then also acquire an influence on our world-view. (1915/1979, p. 153)

[24] There seems to be an exception, namely the entry "Realism, critical", in the second edition of the *Allgemeine Erkenntnislehre* (1925/1974, p. 409). However, as one might expect, the index was not created by Schlick himself, but by his student Feigl.

It is not difficult to see that Schlick is accounting here for the *critical component* of critical realism (as well as of critical idealism!). More precisely, Schlick ties Kantian critical method to the *factual development of science*. Consequently, as Michael Friedman correctly notes, "Schlick aimed to do for Einstein's physics what Kant had done for Newton's, namely, to explain and exhibit the special features of this physics that make it a model or paradigm of coherent rational knowledge of nature" (2001, p. 14). Schlick, it might be added, can therefore be regarded as a follower of Kantian methodology in its relation – and updated application – to science and scientific theory construction.

As for the *realist component*, Schlick seems to aim to continue the Kantian heritage as well. In an article on "Appearance and Essence" – published, typically enough, in the *Kant-Studien* – Schlick declares: "[T]he only natural continuation of Kant's theory of knowledge, to which his system points from various angles, lies not in the idealist but the realist direction, and we arrive at it by a revision of Kant's utterances about the so-called thing-in-itself and its knowability." (1919/1979, p. 282)

The aim to realistically revise the original Kantian doctrine by arguing for the knowability of things-in-themselves could hardly be formulated clearer! Yet, as we have seen before, this realistic strategy deviates rather drastically from what Kant himself in fact intended. Exegetical issues like this notwithstanding,[25] it should be further noted that Schlick affirmatively relied on the critical realists' downgrading of the Kantian faculty of intuition. Or, in his own words:

> Kant has uncritically presupposed that in order to know an object, an *intuition* of the object is ultimately in some way necessary. In the very first sentence of the transcendental aesthetics he says this with complete clarity.[26] But in truth intuition gives us no knowledge whatever; it is wholly inessential for this purpose. It provides, to be sure, an *acquaintance* with objects, but never a knowledge of them. (*Ibid.*)

In consequence of this downgrading of intuition, Schlick arrives at a conception of purely conceptual knowledge that bears obvious similarities with the one propounded by Külpe. Thus, in a chapter titled "Knowing by Means of Concepts", Schlick, in his *Allgemeine Erkenntnislehre*, argues that

> [e]pistemologically, the import of the conceptual function consists precisely in *signifying* or *designating*. Here, however, to signify means nothing more than to *coordinate* or *associate* (*Zuordnen*), that is, to place in a one-one or at most a many-one correspondence ("*Zuordnung*"). To say that objects fall under a certain concept is to say only that we have coordinated or associated them with this concept. (1925/1974, p. 23)

In the footnote attached to this passage, Schlick refers the reader to Külpe's aforementioned theory of concepts as "fixed coordinations", as it can be found

[25] Schlick himself refers the reader to Benno Erdmann as a "competent authority" (1919/1979, p. 282) in interpreting Kant along realistic lines. See, in this connection, esp. (Erdmann 1917).

[26] Compare Kant: "In whatever way and through whatever means a cognition may relate to objects, that through which it relates immediately to them, and at which all thought as a means is directed as an end, is intuition." (B 33)

in volume 1 of *Die Realisierung* (1912, p. 226).[27] Moreover, Schlick approved of Külpe's distinction between "positing" and "determination" of the real, thereby committing himself to an epistemologically affirmative understanding of things-in-themselves (1925/1974, p. 175).

However, Schlick did not unreservedly align himself with Külpe. Rather, he was fairly critical of Külpe's "dualistic world-view". More concretely, Schlick conceded that there are things that can be immediately perceived and things that are "transcendent" (*id.*, p. 238). But he did not go as far as to endorse the view that there are, as it were, *two worlds*, i.e., the world of appearances and the world of things-in-themselves.[28] Rather, Schlick held the whole distinction to be flawed. Accordingly, he wrote:

> There is only *one* reality. And whatever lies within its domain is in principle equally accessible, in its being as well as in its essence, to our cognition. Only a small part of this reality is ever given to us. The remainder is not given. But the separation thus effectuated between the subjective and the objective is accidental in character. It is not fundamental, as the separation between essence and appearance is supposed to be – a separation that we have recognized as not feasible. (*id.*, p. 244)

In short, Schlick committed himself to a certain variant of *monism*.[29]

For reasons of space, Schlick's critique of Külpe cannot be discussed in detail here. But one additional remark concerning the "ontology of science" is worth making. According to Schlick, the critical realists', and, especially, Külpe's substantialist interpretation of things-in-themselves is an anachronism. We simply do not have enough evidence to say that Schlick was directly "influenced" by Cassirer in this respect; but his plea for an essentially *relationalist* understanding of

---

[27] As Thomas Ryckman (1991, pp. 58–61) has correctly pointed out, Schlick's conception of knowledge by coordination has one of its roots in nineteenth-century mathematics. Especially the work of Richard Dedekind, more precisely his account of function and number, is relevant in this connection (cf. Schlick 1925/1974, p. 383). Yet, Külpe's – psychology-based – contribution is, in my view, at least of equal importance as regards the Schlickian understanding of purely conceptual knowledge. For further details, see (Neuber 2012, p. 75, fn. 91).

[28] Compare Schlick [: "The fact that there are real things, some of which are given and some not given, may indeed justify us in distinguishing two classes of real things, but not in assuming two different kinds or levels of reality. Also, Külpe's terminology allows the positing of an unconscious mental reality to seem more natural than is in fact justified, for it permits us, for example, to speak of sensations that are real (= *real*) but at the same time are not also actual (= *wirklich*). [. . .] [T]here is no set of facts that either forces or justifies such a counterposing of two irreducible realities, of which one rests entirely on itself and the other is dependent on the first." (1925/1974, pp. 238–239)

[29] Thus, as early as 1913, Schlick articulated the opinion that "if the concept of knowledge is correctly taken [. . .], we ought in a certain sense [. . .] to regard *any* knowledge that is not purely formal as a knowledge of 'things-in-themselves'" (1913/1979, p. 149). As I learned from Sean Crawford, the source of Schlick's commitment to monism might be located in the *mind-body* problem. This assessment gets supported by Schlick's remark – added in the preface to the second edition of the *Allgemeine Erkenntnislehre* – that, for him, the mind-body problem has "a quite special systematic importance" (1925/1974, p. xii). For further details, see (Feigl 1971, 1975), (Heidelberger 2007).

science comes rather close to Cassirer's "anti-substantialism".[30] Referring himself to the older – empiricist (i.e. Humean and Machian) – critiques of the substantialist point of view, Schlick states that

> an atom or an electron is to be conceived of as a union of qualities that are bound together by definite laws, and not as a substantial *thing*, which bears its qualities as properties and can thus be distinguished from them as their bearer. [...] We need not concern ourselves any further with this idea. [...] In the last analysis, all knowledge is a matter of relations and dependencies, not of things or substances. (*id.*, p. 285)

As Rae Langton (commenting on Guyer 1987, pp. 351–352) has observed, such a relationalist point of view is suitable to provide "a quick route to idealism" (Langton 1998, p. 95). Still, on Schlick's account, *relations have the status of things-in-themselves* and must therefore be interpreted in a realistic way. Schlick writes:

> [A]n object is always a complex of relations. These relations, on Kant's theory, are not immediately given, but must be charged to the account of thought, judgments and concepts. According to the Criticist view, therefore, relations originate in judgments, whereas according to our concept of knowledge judgments are simply correlated with the relations, which exist outside of this correlation. (1925/1974, p. 360)

This is a clear statement in favor of a realist ontology of relations. At the same time it is, because of its being rooted in ontology, a repudiation of a purely epistemological interpretation in the guise of Cassirer (and the Marburg school in general). Thus, it can be said that Schlick combined Cassirerian relationalism with the critical realists' insistence on the independent reality – and knowability! – of things-in-themselves.

Let us conclude by briefly considering a well-known case in point. In his *Space and Time in Contemporary Physics* (1917–1922) Schlick applied his realist-relationalist account of science and nature to Einstein's theory of relativity (cf. Friedman 1999; Howard 1999; Neuber 2012, pp. 102–130).[31] Demarcating his own conception from other, especially positivist interpretations of Einstein's theory, Schlick (by applying his "method of coincidences") argued for a thoroughly realistic understanding of spatiotemporal relations. More generally, Schlick proposed to conceive of the objects of science as mind-independent, explanatory realities:

> There is no argument whatsoever to force us to state that only the intuitional elements, colours, tones, etc., exist in the world. We might just as well assume that elements or qualities which cannot be directly experienced also exist. For example, electric forces can just as well signify elements of reality as colours and tones. They are *measurable*, and there is no reason why epistemology should reject the criterion of reality which is used in physics [...]. The conception of an electron or an atom would then not necessarily be a mere working hypothesis, a condensed fiction, but could equally well designate a

[30] I must admit that I underestimated Cassirer's possible impact on the early Schlick's relationalism in my (Neuber 2012). Thanks to Marco Giovanelli for instructive discussions concerning this point!

[31] One must know in this connection that the manuscript of the *Allgemeine Erkenntnislehre* was already finished in 1916, so that it really makes sense to speak of an "application" to relativity theory here. For further details, see (Engler and Neuber 2006, pp. 36–37) and (Wendel and Engler 2009, p. 75).

> real connection or complex of such objective elements [. . .]. The picture of the world, as presented by physics, would then be a system of symbols arranged into a four-dimensional scheme, by means of which we get our knowledge of reality; that is, *more* than a mere auxiliary conception, allowing us to find our way through given intuitional elements. (1917–1922/1979, p. 265)

From here it is only a short step to the more recent attempts to establish a *structural realist* account of science and nature (cf. Worrall 1989; Ladyman 1998; French and Ladyman 2003; Esfeld and Lam 2006). Given Schlick's relationalist dissolution of (the traditional view of) substantial things, and given further his account of scientific objects as nonetheless real, the conclusion to be drawn is that the early Schlick's reception of the critical realist tradition paved the way for our current discussion concerning structural realism.[32] His interpretation of spatiotemporal relations in the context of Einstein's general theory of relativity might be seen as his paradigmatic contribution to the advancement of the structural realist point of view.

4.1 Coda

This is not the place for an extended discussion of Schlick's alleged Viennese conversion from realism to positivism. However, the major reasons for accepting Feigl's diagnosis of such a conversion should at least be indicated:

- For the early Schlick, relational things-in-themselves exist mind-independently, whereas for the Viennese Schlick things-in-themselves are equivalent with (Russellian) "logical constructions" (cf. Schlick 1926/1979, p. 104).
- For the early Schlick, atoms (and other theoretical entities) are "transcendent", whereas for the Viennese Schlick they form part of Kantian "empirical reality", i.e. of the realm of appearances (cf. Schlick 1932/1979, p. 278).
- For the early Schlick, the statement of a transcendent reality is meaningful and scientifically relevant, whereas for the Viennese Schlick it is, as for Carnap (cf. Carnap 1928), nothing but a "certain state of feeling" (Schlick 1932/1979, p. 281).
- For the early Schlick, realism and positivism exclude each other, whereas for the Viennese Schlick both positions are "not opposed" (*id.*, p. 283).

These are, in my view, differences that must be taken seriously: they must and will be elucidated in a later work (see Neuber, forthcoming).[33]

[32] For similar assessments, see (Bartels 1997), (Friedman 1999, p. 20), (Gower 2000).

[33] I wish to thank Joseph J. Kominkiewicz who constructively commented on an earlier draft of this paper.

References

Allison, H. 1983. *Kant's transcendental idealism: An interpretation and defense*. New Haven: Yale University Press.

Bartels, A. 1997. Die Auflösung der Dinge. Schlick und Cassirer über wissenschaftliche Erkenntnis und Relativitätstheorie. In *Philosophie und Wissenschaften: Formen und Prozesse ihrer Interaktion*, ed. H.-J. Sandkühler, 193–210. Frankfurt a.M: Verlag Peter Lang.

Becher, E. 1914. *Naturphilosophie*. Leipzig: Teubner.

Becher, E. 1915. *Weltgebäude, Weltgesetze, Weltentwicklung: Ein Bild der unbelebten Natur*. Berlin: Reimer.

Boyd, R. 1983. On the current status of the issue of scientific realism. *Erkenntnis* 19: 45–90.

Carnap, R. 1928. *Scheinprobleme in der Philosophie*. Berlin: Weltkreis-Verlag.

Cassirer, E. 1910. *Substanzbegriff und Funktionsbegriff: Untersuchungen über die Grundfragen der Erkenntniskritik*. Berlin: Bruno Cassirer Verlag.

Cassirer, E. 1921. *Zur Einsteinschen Relativitätstheorie: Erkenntnistheoretische Betrachtungen*. Berlin: Bruno Cassirer Verlag.

Cohen, H. 1902. *Logik der reinen Erkenntnis*. Berlin: Bruno Cassirer Verlag.

Drake, D., A. Lovejoy, J.B. Pratt, A. Rogers, G. Santayana, R.W. Sellars, and C.A. Strong. 1920. *Essays in critical realism: A co-operative study of the problem of knowledge*. London: Macmillan.

Engler, F.O., and M. Neuber (eds.). 2006. *Moritz Schlick – Kritische Gesamtausgabe, Abteilung I, Band 2*. Wien/New York: Springer.

Erdmann, B. 1917. *Die Idee von Kants Kritik der reinen Vernunft: Eine historische Untersuchung*. Berlin: Reimer.

Esfeld, M., and V. Lam. 2006. Moderate structural realism about space-time. *Synthese* 160: 27–46.

Feigl, H. 1971. Some crucial issues of mind-body monism. *Synthese* 22: 295–312.

Feigl, H. 1975. Russell and Schlick: A remarkable agreement on a monistic solution of the mind-body problem. *Erkenntnis* 9: 11–34.

Feigl, H. [1963] 1981. The power of positivistic thinking. In *Inquiries and provocations: Selected writings, 1929–1974*, 38–56. Dordrecht/Boston/London: Reidel.

French, S., and J. Ladyman. 2003. Remodelling structural realism: Quantum physics and the metaphysics of structure. *Synthese* 136: 31–56.

Friedl, J., and H. Rutte (eds.). 2008. *Moritz Schlick – Kritische Gesamtausgabe, Abteilung I, Band 6*. Wien/New York: Springer.

Friedman, M. 1999. *Reconsidering logical positivism*. Cambridge: Cambridge University Press.

Friedman, M. 2000. *A parting of the ways: Carnap, Cassirer, and Heidegger*. Chicago/La Salle: Open Court.

Friedman, M. 2001. *Dynamics of reason: The 1999 Kant Lectures at Stanford University*. Stanford: CSLI Publications.

Friedman, M. 2005. Ernst Cassirer and the philosophy of science. In *Continental philosophy of science*, ed. G. Gutting, 71–84. London: Wiley-Blackwell.

Friedman, M. 2012. Scientific philosophy from Helmholtz to Carnap and Quine. In *Rudolf Carnap and the legacy of logical empiricism*, ed. R. Creath. Dordrecht/Heidelberg/New York/London: Springer.

Geymonat, L. 1985. Entwicklung und Kontinuität im Denken Schlicks. In *Zurück zu Schlick: Eine Neubewertung von Werk und Wirkung*, ed. B. McGuiness, 24–31. Wien: Hölder-Pichler-Tempsky.

Gower, B. 2000. Cassirer, Schlick and 'structural' realism: The philosophy of the exact sciences and the background to early logical empiricism. *British Journal for the History of Philosophy* 8: 71–106.

Guyer, P. 1987. *Kant and the claims of knowledge*. Cambridge: Cambridge University Press.

Heidelberger, M. 2007. From Neo-Kantianism to critical realism: Space and the mind-body problem in Riehl and Schlick. *Perspectives of Science* 15: 26–48.

Henckmann, W. 1997. Külpes Konzept der Realisierung. *Brentano-Studien* 7: 197–208.
Howard, D. 1999. Point coincidences and pointer coincidences: Einstein on invariant structure in spacetime theories. In *The history of general relativity IV: The expanding worlds of general relativity*, ed. H. Goenner, J. Renn, J. Ritter, and T. Sauer, 463–500. Boston: Birkhäuser.
Kant, I. [1781/1787] 1998. *Critique of pure reason*. Translated and edited by P. Guyer and A.W. Wood. Cambridge: Cambridge University Press.
Külpe, O. 1893. *Grundriss der Psychologie: auf experimenteller Grundlage dargestellt*. Leipzig: Engelmann.
Külpe, O. 1910. *Einleitung in die Psychologie*. Leipzig: Hirzel.
Külpe, O. 1912. *Die Realisierung: Ein Beitrag zur Grundlegung der Realwissenschaften*, vol. 1. Leipzig: Hirzel.
Külpe, O. 1920. *Die Realisierung: Ein Beitrag zur Grundlegung der Realwissenschaften*, vol. 2. Leipzig: Hirzel.
Külpe, O. 1923. *Die Realisierung: Ein Beitrag zur Grundlegung der Realwissenschaften*, vol. 3. Leipzig: Hirzel.
Ladyman, J. 1998. What is structural realism? *Studies in History and Philosophy of Science Part A* 29: 409–424.
Langton, R. 1998. *Kantian humility: Our ignorance of things in themselves*. Oxford: Clarendon.
Lindenfeld, D. 1978. Külpe and the Würzburg School. *Journal of the History of the Behavioral Sciences* 14: 132–141.
Natorp, P. 1910. *Die logischen Grundlagen der exakten Wissenschaften*. Leipzig: Teubner.
Neuber, M. 2011. Zwei Formen des transzendentalen Revisionismus: 'Wissenschaftliche Philosophie' beim frühen Ernst Cassirer und beim frühen Moritz Schlick. *Kant-Studien* 102: 455–476.
Neuber, M. 2012. *Die Grenzen des Revisionismus: Schlick, Cassirer und das 'Raumproblem'*. Wien/New York: Springer.
Neuber, M. forthcoming. Schlick und die 'Wende der Philosophie': Vom kritischen Realismus zum logischen Empirismus (und wieder zurück?), in: Matthias Neuber (ed.), Husserl, Cassirer, Schlick. 'Wissenschaftliche Philosophie' im Spannungsfeld von Phänomenologie, Neukantianismus und logischem Empirismus. Dordrecht/Heidelberg/New York/London: Springer.
Paton, H. 1936. *Kant's metaphysics of experience*. London: G. Allen and Unwin.
Prauss, G. 1974. *Kant und das Problem der Dinge an sich*. Bonn: Bouvier.
Psillos, S. 1999. *Scientific realism: How science tracks truth*. London: Routledge.
Richardson, A. 1997. Toward a history of scientific philosophy. *Perspectives on Science* 5: 418–451.
Richardson, A. 2008. Scientific philosophy as a topic for history of science. *Isis* 99: 88–96.
Riehl, A. 1876/1908. *Der philosophische Kritizismus: Geschichte und System*, vol. 1. Leipzig: Engelmann.
Riehl, A. 1879/1925. *Der philosophische Kritizismus: Geschichte und System*, vol. 2. Leipzig: Kröner.
Riehl, A. 1883. *Ueber wissenschaftliche und nichtwissenschaftliche Philosophie*. Berlin.
Riehl, A. 1887/1926. *Der philosophische Kritizismus: Geschichte und System*, vol. 3. Leipzig: Kröner.
Russell, B. 1910–1911. Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society (New Series)* XI: 108–128.
Ryckman, T. 1991. *Conditio Sine Qua Non: Zuordnung* in the early epistemologies of Cassirer and Schlick. *Synthese* 88: 57–95.
Sankey, H. 2008. *Scientific realism and the rationality of science*. Aldershot: Ashgate.
Schlick, M. 1913/1979. Is there intuitive knowledge? In *Moritz Schlick: Philosophical papers, vol. 1 (1909–1922)*, ed. H. Mulder and B. van de Velde-Schlick, 141–152. Dordrecht/Boston/London: Reidel.
Schlick, M. 1915/1979. The philosophical significance of the principle of relativity. In *Moritz Schlick: Philosophical papers, vol.1 (1909–1922)*, ed. H. Mulder and B. van de Velde-Schlick, 153–189. Dordrecht/Boston/London: Reidel.

Schlick, M. 1917–1922/1979. Space and time in contemporary physics: An introduction to the theory of relativity and gravitation. In *Moritz Schlick: Philosophical papers, vol. 1 (1909–1922)*, ed. H. Mulder and B. van de Velde-Schlick, 207–269. Dordrecht/Boston/London: Reidel.

Schlick, M. 1919/1979. Appearance and essence. In *Moritz Schlick: Philosophical papers, vol. 1 (1909–1922)*, ed. H. Mulder and B. van de Velde-Schlick, 270–287. Dordrecht/Boston/London: Reidel.

Schlick, M. 1925/1974. *General theory of knowledge*. Trans: A.E. Blumberg. Wien/New York: Springer.

Schlick, M. 1926/1979. Experience, cognition and metaphysics. In *Moritz Schlick: Philosophical papers, vol. 2 (1925–1936)*, ed. H. Mulder and B. van de Velde-Schlick, 99–111. Dordrecht/Boston/London: Reidel.

Schlick, M. 1932/1979. Positivism and realism. In *Moritz Schlick: Philosophical papers, vol. 2 (1925–1936)*, ed. H. Mulder and B. van de Velde-Schlick, 259–284. Dordrecht/Boston/London: Reidel.

Schnädelbach, H. 1983. *Philosophie in Deutschland 1831–1933*. Frankfurt a.M: Suhrkamp.

Sellars, R.W. 1969. *Reflections on American philosophy from within*. Notred Dame: University of Norte Dame Press.

Strawson, P. 1966. *The bounds of sense: An essay on Kant's critique of pure reason*. London/New York: Routledge.

Vaihinger, H. 1881/1892. *Commentar zu Kants Kritik der reinen Vernunft: Zum 100jährigen Jubiläum desselben herausgegeben*, 2 vol. Stuttgart: Spemann.

Wendel, H.-J., and F.-O. Engler (eds.). 2009. *Moritz Schlick – Kritische Gesamtausgabe, Abteilung I, Band 1*. Wien/New York: Springer.

Willaschek, M. 2001. Die Mehrdeutigkeit der kantischen Unterscheidung zwischen Dingen an sich und Erscheinungen: Zur Debatte um Zwei-Aspekte- und Zwei-Welten-Interpretationen des transzendentalen Idealismus. In *Kant und die Berliner Aufklärung: Akten des IX. Internationalen Kant-Kongresses. Band II: Sektionen I-V*, ed. V. Gerhardt, R.-P. Horstmann, R. Schumacher, V. Gerhardt, R.-P. Horstmann, and R. Schumacher, 678–690. Berlin/New York: de Gruyter.

Worrall, J. 1989. Structural realism: The best of both worlds? *Dialectica* 43: 99–124.

Wundt, W. [1896] 1910. Über naiven und kritischen Realismus. In *Kleine Schriften*, vol. 1, 259–510. Leipzig: Engelmann.

Realism Without Mirrors

Henrik Rydenfelt

1 Introduction

Contemporary pragmatists, especially those who in one way or another follow Richard Rorty's lead, have contested the philosophical paradigm they have referred to as "representationalism": the idea that our claims, thoughts, beliefs and the like describe, reflect or are "about" the world or reality. These "new pragmatists" have often been seen to be in an antagonistic relationship with their antecedents. The pragmatists of the turn of the twentieth century, Charles S. Peirce, William James and John Dewey, all advanced at least moderate forms of realism, towards which the non-representationalist position is taken to be hostile. Contrary to this common assumption, it is my aim to show that, on the one hand, the early pragmatists could adopt the basic tenets of non-representationalism and that, on the other hand, the form of realism they developed could complement the views of their contemporary namesakes. The result of this novel combination is a realism which does without representationalism.

In what follows, I will first describe meta-ethical expressivism and its current non-representationalist offspring, the *global expressivist* view that Huw Price defends in the papers collected in *Naturalism without Mirrors* (2011b) and several other writings. I will then proceed to argue that the views of the pragmatists allow for an expressivist and non-representationalist interpretation. The assumption that this cannot be the case is based on the received wisdom that realism and non-representationalism are incompatible philosophical views – an assumption I will contest by arguing that realism, conceived of as an ontological (as opposed to a semantic) view, is fully compatible with non-representationalism. Moreover,

H. Rydenfelt (✉)
Department of Social Research, University of Helsinki, Snellmaninkatu 14 B, Helsinki 00014, Finland
e-mail: henrik.rydenfelt@helsinki.fi

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_47,

I will propose that Charles S. Peirce's account of the scientific method and its realist underpinnings – the view which I will call *hypothetical realism* – is not derived from representationalist considerations but, rather, can be sustained within a non-representationalist framework. Finally, as an example of how the novel view here developed can help us to reconceptualize realism in different domains, I will consider its extension to normative (or moral) realism.

2 Varieties of Expressivism

A central debate in meta-ethics of the past decades was first conceived of as one between cognitivism and non-cognitivism. Traditional non-cognitivism, originally proposed by thinkers such as Stevenson (1944) and Ayer (1952), held that moral (or more broadly normative) statements do not express beliefs but, rather, non-cognitive states such as emotions and desires. As such, normative statements – unlike non-normative ones – were argued to have no truth-values. Cognitivism, in turn, is the traditional view that normative statements – like non-normative statements – describe the world, express beliefs, and have truth-values like any other statement.

Non-cognitivism fell out of philosophical favor by the late 1960s due to criticism by Peter Geach and John Searle, who argued that the non-cognitivist has no plausible account of how statements expressing non-cognitive attitudes enter into logical relations such as those involved in deductive inferences. For a while this Frege-Geach-Searle objection was held to be a decisive refutation of non-cognitivism. Since the 1980s, however, philosophers working under the *expressivist* banner have attempted to tackle this problem in various ways. Simon Blackburn's (1988, 1998) expressivism set out to earn the right for a notion of truth for normative claims, non-cognitivistically understood, by combining non-cognitivism with a deflationary account of truth. This quasi-realist approach was to make sense of the realist-seeming nature of such claims while retaining a crucial contrast between normative and non-normative statements. At the same time, expressivist views became more encompassing. In Allan Gibbard's (1990, 2003) expressivist understanding of language, non-normative statements themselves are conceived of as *expressions* of belief-like states instead of *descriptions* of the world. This approach, Gibbard has argued, will ultimately enable the expressivist to cast the Fregean concerns of Geach and Searle behind.

While whether Gibbard is correct is a question far beyond the scope of this discussion, as a consequence of these developments, the original debate between non-cognitivism and cognitivism was now conceived of as one between two different approaches to *language*: expressivism and (what is now often called) descriptivism. This spreading of the expressivist stance beyond its initial domain of normative language has paved the way for Huw Price's (2011a,b) *global expressivism*. Price, to an extent following Richard Rorty's lead, contests the

traditional philosophical notion of representation – the idea that our thoughts, claims, statements and the like are "about" or describe some "facts". In his view – which he has likened to Robert Brandom's (1994, 2000) inferentialism – claims rather express our functional, behavioral and inferential stances or commitments. When making such commitments explicit in a discourse with others, our claims attain their typical assertoric shape and propositional form; Brandom's view of assertion as making inferential commitments explicit is one account of how this takes place.

People working under the expressivist banner are thus variously divided. Blackburn has retained a descriptivist view about non-normative statements, which his quasi-realism is not intended to cover. His expressivism is thus *local*. In turn, Gibbard and more recently Mark Schroeder (2008) have extended expressivism to non-normative language. In their view, however, non-normative judgments gain their truth-conditions, or propositional content, from the beliefs that they express: for example, the statement "a cat is on a mat" expresses the belief *that* a cat is on a mat, and it is the propositional content of this belief (i.e. that a cat is on a mat) which then forms the truth-conditions of that statement. While in a sense encompassing both normative and non-normative uses of language, this approach thus splits language into (at least two) distinct regions, one of which still involves a robust notion of representational content. Such expressivism is, we might put it, *regional*.

The *global* expressivism advanced by Price (and Brandom, in Price's reading) adopts a crucially different perspective. Eschewing any robust concept of representation, it does without any contrast with between normative and non-normative statements (thoughts, beliefs) in representationalist terms. The differences between these sort of thoughts or commitments are functional rather than representational by nature. Price's global expressivism is thus a (globally) *non-representationalist* position. It maintains that claims or statements express commitments of a practical sort. Such commitments do not involve a "representation" of reality but are functional or behavioral in nature.[1] While nothing prevents the non-representationalist from employing standard philosophical notions such as "content" and "proposition" (or, indeed, "representation"), these notions are to be construed in a manner that does not involve a robust representational relation from claims, thoughts and mental states to reality. Accordingly, truth is not to be understood as "correspondence" of a thought or a claim with reality but as a linguistic or grammatical device as a variety of deflationary or minimalist accounts have maintained. This non-representationalist approach is currently gaining ground under the label of *pragmatism*.

[1]To be precise, in order to avoid potentially paradoxical-seeming statements, Price's global expressivist nowhere *denies* that our claims do not "represent" the world. Rather, the global expressivist *avoids* making such claims in giving her account of language.

3 Pragmatism and Expressivism

In what follows, I will argue that pragmatism in its *classical* form can be interpreted as combining realism – in particular, a form of scientific realism as proposed by Charles S. Peirce – with a non-representationalist position closely akin to that advanced by their contemporary offspring. As the very thought that the pragmatists (as I will refer to its classics, especially Peirce) could have approved of anything like the non-representationalist view is rather contentious, I will offer three different considerations in defence of the first claim. Firstly, the pragmatists anticipated the local expressivist view pertaining to moral language. Secondly, the very starting point of pragmatism, the *pragmatist maxim*, implies that our claims are primarily to be taken to express our functional or practical dispositions and commitments – the shared starting point of regional and global expressivism. And thirdly, the form of realism that the pragmatists advanced is not only independent of representationalist assumptions but moreover gives grounds to a non-representationalist (or global expressivist) interpretation of pragmatism; indeed, in the following Sections I will argue that if Peirce had been a proponent of a straightforward representationalist picture, his defence of scientific realism would *not* have taken the shape it did.

According to my first claim, the pragmatists advanced a view of moral language which bears resemblance to contemporary (non-cognitivist) expressivism. Obviously, such a claim is bound to be somewhat anachronistic. It would not do much injustice to say that the contemporary meta-ethical debate largely begins with G. E. Moore's (1903) famous Open Question Argument, which challenges the cognitivists to make good sense of what sort of properties normative terms such as "good" and "right" predicate. Peirce, James and Dewey never took up Moore's argument, and their writings do not involve much by way of sustained discussions on meta-ethical topics.[2]

Considering James's usual lack of interest in the systematic development of themes in the philosophy of language, it may come as something of a surprise that their probably most sustained statement on normative thought and language appears in his relatively early address, "The Moral Philosopher and the Moral Life" (1891). There James maintains that moral ideals are dependent on the desires or demands of "sentient beings" such as ourselves:

> Physical facts simply *are* or are *not*; and neither when present or absent, can they be supposed to make demands. If they do, they can only do so by having desires; and then they have ceased to be purely physical facts, and have become facts of conscious sensibility.

[2]Common to the pragmatists is the view that our moral ideas are open to revision quite like our other ideas. At least initially, this could be taken to imply a cognitivist position according to which moral thoughts and claims are responsive to some (moral) facts. My ultimate proposal here also attempts to make sense of the idea that normative thought is open to revision in a manner analogous to non-normative thought. If successful, then, the position here developed will also account for the cognitivist-seeming views of the pragmatists.

> Goodness, badness, and obligation must be realized somewhere in order really to exist; and the first step in ethical philosophy is to see that no merely inorganic 'nature of things' can realize them. (James 1897, p. 190)

From this, James draws the following conclusion concerning moral language:

> [T]he words 'good,' 'bad,' and 'obligation' [. . .] mean no absolute natures, independent of personal support. They are objects of feeling and desire, which have no foothold or anchorage in Being, apart from the existence of actually living minds. (James 1897, p. 197)

Put in contemporary terms, James here argues that moral terms – unlike claims concerning physical facts – do not refer to properties out there in the world (or "absolute natures") but rather give expression to our conative states of mind (or the "objects of feeling and desire"). This picture obviously bears resemblance to Simon Blackburn's *local* expressivism.

According to my second proposal, there is reason to think that the pragmatists could accept the extension of the expressivist approach beyond its traditional moral scope. The origin of pragmatism is in the pragmatist maxim formulated by Charles S. Peirce (1878) and later advanced in a somewhat different version by William James (1907, ch. 2), who was the first to use the term "pragmatism" in print, giving full credit to Peirce, in 1898. This maxim is sometimes glossed as the pragmatist account of meaning; however, this is a rather crude simplification. In Peirce's original formulation, pragmatism enables us to grasp a *dimension* of meaning aside from acquaintance-based familiarity and definitional understanding of a concept. To attain a further, or "third" grade of clarity of our claims (or the concepts they involve), we are to consider the imaginable practical consequences differences in the conduct of an agent who believes that claim (or a claim which entails that concept): to find out what is (in this sense) meant by "hard", we are to consider the conduct of someone who believes that something is hard.[3] Moreover, any meaningful claim is one that would make a practical difference to the conduct of an acting agent. The pragmatist maxim relies on the contention that beliefs by their nature involve a preparedness to act under some conceivable circumstances: they are *habits* (although ones that never may bear fruit in actual conduct). As with regional and global expressivism, pragmatism hinges on the idea that our claims – including our non-normative claims – express functional or dispositional states of a practical nature.

This starting point however leaves open the issue of the role of representation in the pragmatist account. Regional expressivism, as we saw, still retains a representationalist order of explanation: it views beliefs as attitudes towards propositions (or "representations"), the acceptance of which will then involve some practical consequences to the believing agent's conduct. Global expressivism or non-representationalism implies a reverse order of explanation. It does not begin

[3]Pragmatists often connect these consequences to the conduct of an agent to the expectations of what will occur in experience, if the accepted idea or claim is true (cf. Rydenfelt 2009a). However, in my view even this connection assumes the perspective of the Peircean scientific method, to which I will presently return.

with a received account (propositional) *content*, and with beliefs as attitudes towards such content. Instead, it sets out with the notion of beliefs as functional or dispositional states. Propositional content, in turn, is considered in terms of functional and inferential commitments that are put forward in our assertoric practices. Semantic notions such as "proposition", "content" and "representation", are technical and philosophical devices for accounting for what is being believed, thought or said – for what, for example, is common to different verbal manifestations of the same claim or thought.

As we may well expect, Peirce offers no single account pertaining to the role of representation (understood in this contemporary fashion). At points, he holds on to the traditional notion of beliefs as attitudes towards propositions, lending to a representationalist interpretation (e.g. Peirce 1903, p. 139). But many of his discussions on the nature of beliefs do not at all invoke representationalist notions at all. Crucially, Peirce nowhere appears to maintain that the practical operation of beliefs as habits is due to a specific "representing" function, or their being "about" some realities, nor that believing requires awareness of a (representational) content or proposition. (Indeed, when we ascribe beliefs to animals, such as dogs, we cannot even expect them to be able to formulate the content of their belief in any such manner.) The basic ideas of the non-representationalist view could thus be accepted by Peirce; as I will proceed to argue in what follows, this view is better suited to Peirce's discussion of truth and the scientific method than the representationalist alternative.

It is not my intention to claim that the pragmatist account of the nature of claims and the sort of functional states they express is in every detail the same as that advanced by their contemporary namesakes. Robert Brandom has noted that the classical pragmatists' notion of meaning centres on practical consequences in conduct. This is in distinction to his own account by which the meaning of a claim (or its conceptual content) is composed of both its consequences in action (or its "exit rules") *and* the circumstances where one becomes entitled or committed to endorse that claim (or its "entry rules"). Brandom turns this fact into a criticism of pragmatism, which in his view leads to a semantic theory which is "literally one-sided" as it identifies "propositional contents exclusively with the consequences of endorsing a claim, looking downstream to the claim's role as a premise in practical reasoning and ignoring its proper antecedents upstream" (2000, pp. 64, 66).

I'm mostly in agreement with Brandom's analysis: there is a difference of this sort – at least one of emphasis – between his view and that of the early pragmatists. However, I do not think this should be taken to imply that the latter account is crucially limited; instead, the pragmatist view comes with a distinct advantage.[4] Consider, for example, an atheist and a religious fundamentalist, who not only sharply disagree about whether there is a God but have radically different notions of

[4]In addition, I do not intend to endorse the whole of Brandom's reading by which the classical pragmatists advanced an instrumentalist account of truth in terms of the success of practices. For a useful critical discussion, see Pihlström (2007, pp. 272–275).

what counts as evidence for or against such a claim. It is a well-known consequence of Brandom's account that "God" and "God exists" then mean different things for them: they have differing entry rules for the commitment expressed by that claim.

Brandom, of course, is not alone. Similar considerations have led many to think that the God-talk of the atheist and the fundamentalist belong to different language games altogether. But such a view renders hopeless any effort to find common ground for settling issues between such interlocutors. Pragmatism in its classical version makes things easier for us. It allows us to say that, despite having differing conceptions of what makes for good evidence for their beliefs, the atheist and the fundamentalist talk about the *same* thing insofar as the belief in God (or lack thereof) would similarly affect their conduct. By not making conceptions of evidence integral parts of the meanings of our claims, the pragmatist view leaves open the issue of how to settle such conflicts of opinion. As we will shortly see, this difference between the classical pragmatists and their contemporary followers is one that makes a difference.

4 A Compatibility Claim

So far I have argued that the classical pragmatists, especially Peirce, could accept the basic tenets of the global expressivist and non-representationalist interpretation. This however goes against the standard contemporary accounts of their views. Proponents of classical pragmatism, especially the Deweyans of today, have tended to criticize Rorty for reading the classics, especially Dewey, as non-representationalists (e.g. Rorty 1982). The most important reason for this criticism appears to be their conviction – I think correct – that Dewey was more inclined towards a realist position than Rorty has tended to admit. Where Rorty and the Deweyans *do* however agree is that even a moderately realist reading of Dewey would involve a number of representationalist assumptions. Consequently, Rorty's Dewey is not much of a realist, and the Deweyans' Dewey is a representationalist. Things get even more polarized when Peirce – the indubitable arch-realist – and Rorty are pitched against one another (e.g. Haack 1998). This has lead to the common assumption that there are *two* pragmatisms, the one the realist strand beginning with Peirce, the other the non- or anti-representationalist brand promoted by Rorty and arguably anticipated by James and Dewey. But while there are various differences between the pragmatists, old and new – much too various to be accounted for here – perhaps this issue is not after all such a great divider. My suggestion is that this picture of two distinct traditions is rather dependent on the received wisdom that non-representationalism is incompatible with realism.

An important reason for this latter assumption is the fact that the debate over expressivism has been largely conducted in meta-ethics, where the expressivist position about normative claims has been contrasted with realism on the non-normative side of things. Up till this point it has been taken as a matter of course that expressivism about normative language, understood in the non-cognitivist manner,

results in an ontological position which does without realism. More than that, the expressivist view has often been taken to imply at least some commitment to anti-realism. While Simon Blackburn explicitly denies that his expressivism should be understood as the claim that there are no moral facts or properties, contrary to his intentions, his "projectivism" may easily be taken to imply a form of anti-realism about morality – the view that moral "facts" are merely projected on a reality strictly speaking composed of non-moral facts such as those studied by natural science. Thus, for example, Mark van Roojen's (2004) entry "Moral Cognitivism vs. Non-cognitivism" in the *Stanford Encyclopedia of Philosophy* begins with the statement "Non-cognitivism is a variety of irrealism about ethics with a number of influential variants".

However, it should be noted that expressivism is a view about the nature of claims, thoughts, mental states and the like, not about *what there is* in general terms. As such, it is not an *ontological* position at all but a set of views itself open to a variety of ontological stances and interpretations. This becomes especially evident when expressivism is globalized to the non-representationalist stance. Expressivism appears as an anti-realist or "irrealist" view only when set against some *real* "realism". Losing any such contrast, the non-representationalist view has no particular ontological implications; indeed, we could hardly make sense of what *global* quasi-realism would mean. Global expressivism, or non-representationalism, should be considered an ontologically *neutral* position.

For another example from the direction of the discussions on realism, however, the entry "Realism" by Alexander Miller (2002) in the *Stanford Encyclopedia* lists expressivism as one of the ways in which to "resist" the existence dimension of realism, or the claim that some things exist. The underlying reason for this is that an expressivist interpretation frees claims made in some domain of any commitment to the existence of some facts to make those claims true. But even as the non-representationalist position as presented here rejects such a commitment in a global fashion, as well as involves a denial of a traditional correspondence account of truth, it is hardly evident that realism itself is wedded to any particular *semantic* picture. As Michael Devitt (1991) has for long emphasized, realism conceived of as an ontological position is distinct from any semantic views that we might hold, most centrally a correspondence theory of truth. There is no *prima facie* reason to think that any of the semantic views that non-representationalism sets out to contest are necessary for realism as an *ontological* position.

At the outset, then, non-representationalism and realism are not mutually exclusive, or jointly incoherent views. Obviously, the question of why *be* realist, if one is a non-representationalist, still remains open.[5] Once the burdens of representationalism have been relieved, there may be little temptation to subscribe to an ontological view of any kind. In the hands of the global expressivist, ontology may receive a treatment similar to the minimalist purging of robustness that our

[5]I'm indebted to Jonathan Knowles for discussions on this point.

central semantic terms already have. Perhaps "a cat is on the mat" commits one, minimally, to claims such as "it *is a fact* that a cat is on the mat", or "it *is the case* that a cat is on the mat". But by analogy to the deflationary truth-predicate, the non-representationalist may argue that the italicized phrases don't really add anything ontologically robust to the nature of the commitment.

However, I think there is a story to be told in favour of a form of realism, derivable from the pragmatists – one that is not dependent on a representationalist picture of assertoric activities themselves, the picture prone to be deflated in the expressivist fashion. The source of this commitment to realism is the pragmatist perspective on truth as the aim of inquiry. If anywhere, it is here that the paths of the pragmatists, new and old, begin to diverge.

5 Pragmatists on Truth

In the contemporary philosophical debate over truth, there are two main contenders: the correspondence theory and a variety of deflationary or minimalist accounts. The former maintains that truth is a sort of a fit between a truth-bearer (idea, proposition, belief) and a truth-making reality. This account is often presented as an intuitively plausible *analysis* of our predicate "true". Instead of setting about to uncover the meaning of truth, the latter, deflationary view attempts to give an account of the *use* of the truth predicate in our assertoric practices, an account that the deflationist argues is exhaustive of the predicate itself. As a third alternative, there is a variety of epistemic accounts of truth which attempt to analyze the concept of truth in terms of epistemic notions, such as justification, warrant and belief.

Many have made the mistake of thinking of the pragmatists as attempting to participate in the analytic project. For example, James's elucidations of truth in terms of what works or what would be useful to believe have been used to ridicule the pragmatist position, as if James had aspired to uncover the *conceptual content* of "true". For someone playing the analytic game, it is childishly easy to find counterexamples to any such analysis.[6] In turn, drawing from notions such as *use* and *practices* is what has led many to assimilate the deflationary position with the pragmatist one. However, the pragmatists offered an approach to truth which differs from both of the accounts currently in vogue. Rather than focusing on the conceptual content or the use of the truth predicate, they approached truth in terms of the sort of *beliefs* that we should, or would be better of to have. In James's famous dictum, truth is just the "good in the way of belief". This notion of truth is indistinguishable from their notion of inquiry: truth is the *aim* of inquiry or belief (cf. Rydenfelt 2009b). In one sense, the pragmatist approach is thus deeply epistemic: its notion of truth is that

[6]To an extent, James himself is to be blamed for the confusion. For some reason, he decided to title his 1909 collection of articles on the topic *The Meaning of Truth*.

of the aim of inquiry. But in another sense, this is not the case: as we will presently see, the pragmatist does not maintain that truth is *analyzable* as any such aim.[7]

The central pragmatist text in this respect is Peirce's classic piece, "The Fixation of Belief" (1877). There, Peirce's starting point is the pragmatist notion of inquiry as the move from the unsettling state of doubt to the settlement of opinion, or belief.[8] "Fixation" then discusses four different ways of settling opinion or aims of inquiry, in effect four different notions of truth from the pragmatist perspective. The first of the methods is tenacity, the steadfast clinging to one's opinion. However, under the influence of what Peirce calls the "social impulse", this method is bound to fail. The disagreement of others begins to matter, and the question becomes how to fix beliefs for *everyone* instead of merely for oneself. The three latter methods Peirce discusses are ones attempting to reach such a shared opinion across believers (or inquirers). Contemporary scholars of pragmatism have referred to this demand as underlying the (pragmatist) notion of *objectivity*, or of a standard of opinion beyond one's current views and inclinations (Misak 2000, pp. 3, 52; Short 2007, pp. 324–325).

Interestingly enough, this same phenomenon is reflected in the revised version of the deflationary account of truth propounded by Huw Price. The aspect of our concept of truth as used in our assertoric practices that Price (1998, 2003) has drawn attention to is its function as a "convenient friction" pointing towards a disagreement to be resolved. The response "that's not true," invites disagreement at least in many of our discourses. For contrast, Price envisions a group of "merely-opinionated asserters," whose assertoric practices include a deflationary truth predicate but do not involve this phenomenon. For these speakers, the concept of truth is merely used to register one's agreeing or conflicting opinion, but disagreement will not matter.[9]

The phenomenon Price points towards in practices of assertion is intimately related to the demand for a shared opinion which Peirce detects in our practices of fixing belief. Price's merely-opinionated asserters are akin to Peirce's tenacious

[7]During the past decades, the pragmatist perspective has been sometimes assimilated to the epistemic conception of truth largely due to the influence of Hilary Putnam (1981).

[8]Peirce points out that we might think this is not enough but insist that "we seek, not merely an opinion, but a true opinion". However, this "fancy" is immediately dispelled: "we think each one of our beliefs to be true, and, indeed, it is mere tautology to say so" (1877, p. 115). Here Peirce appears to anticipate contemporary deflationist accounts of truth (cf. Short 2007, pp. 332–333).

[9]The fault of the usual deflationist view, Price argues, is that it does not take into account this aspect of the concept of truth present in our assertoric practices. I am not however sure if the traditional deflationist should be very concerned. At least according to deflationary views which maintain that the concept of truth is a device of disquotation or reassertion, the "friction" phenomenon might only be expected, but not due to some special power invested in our concept of *truth* itself: if disagreement matters, it will matter even if we lacked that concept. The assertion "London is the capital of France", being met with the response, "London is not the capital of France", will invite disagreement just as much (or as little) as it would if the latter speaker had the conceptual capacity of simply pointing out: "that's not true".

believers: they do not aim to coordinate their opinions. Indeed, arguably they are much the same people: the lack of friction of the merely-opinionated speaks to their tenacity, and is derivative thereof. Why disagreement matters in many of our assertoric practices is because we, unlike the tenacious believers, aim to coordinate our underlying commitments.[10]

The question that Peirce addresses – and Price doesn't – is how to resolve such disagreement. The second method Peirce discusses is a straightforward way of doing so: by this method of authority, a power such as that of the state forces a single opinion upon everyone by brute force, even the elimination of dissidents. However, this method ultimately becomes questionable because of the arbitrariness of its results. A "wider sort of social feeling" will show that the opinions dictated by the authority are mostly accidental: different peoples at different ages have held differing views (Peirce 1877, p. 118). The third, *a priori* method attempts to rectify this problem by fixing opinion so that its content would not be arbitrary. Instead, opinion is to be settled, under conditions of liberty, by what is agreeable to reason. However, this method "makes of inquiry something similar to the development of taste; but taste, unfortunately, is always more or less a matter of fashion" (1877, p. 119). The actual development of human opinion will show that this method does not lead to any stable consensus – a result that we will ultimately find unsatisfactory (cf. Rydenfelt, forthcoming).

To avoid the problems of the *a priori* method, it is required to develop a method which does not make our belief dependent of our subjective opinions and tastes altogether, "by which our beliefs may be determined by nothing human, but by some external permanency" (1877, p. 120). This method is the scientific one: it depends on the assumption that there is an independent reality, which "affects, or might affect, every man" (1877, p. 120). Truth, from its point of view, is the opinion which accords with a reality independent of our opinions of it. The hypothesis that underlies the scientific method is that there are things independent of whatever any number of us may think – *hypothetical realism*, as I will call it. This assumption finally makes intelligible the attainment of objectivity, or the possibility of reaching a single answer to any question across inquirers.

The pragmatist view is not an epistemic account of truth which attempts to analyze our concept of truth in terms of epistemic notions, and Peirce nowhere identifies truth as understood within the scientific method with epistemic notions: scientific inquiry is not just *any* investigation that would bring about a consensus among inquirers, but one that has finding out how things are independently of us as its goal. However, epistemic concepts play a central role in the Peircean picture of science. Making the scientific notion of truth more concrete, Peirce suggests that truths are those opinions that would continue to withstand doubt were scientific

[10]The norm of sincerity present in many of our discourses could arguably be due to this fact: we do not merely want others to pay lip service, but need to detect disagreement that ought to be resolved.

inquiry pursued indefinitely (1878, p. 139).[11] Drawing from scientific *practice* prevents Peirce's notion of scientific method from settling on an inexplicable notion of "correspondence".

The pragmatist approach to truth just delineated implies two lessons crucial for the discussion at hand. Firstly, the pragmatist does not draw from conceptual resources in defending the scientific method and the ensuing hypothetical realism. The overall pragmatist approach on truth as the aim of inquiry rather implies that beliefs as such are not to be taken to stand, as if automatically, in any robust semantic relation with "the world". Peirce's discussion of the methods of fixing belief traces the development of the aim of inquiry – the development of the notion of truth, pragmatically conceived. It is not however intended as a method-neutral *argument* for the scientific method. Indeed, if Peirce relied on the idea that beliefs (or claims) "represent" an independent reality and it is *hence* that the scientific method is successful, his discussion of the different methods of fixing belief would be moot: science would win as if by default. If anything, the converse is the case. The fact that the choice between the methods is a genuine one shows that, for Peirce, realism does not simply fall out of a representationalist picture. The scientific method as the product of a substantial development is independent of any representationalist assumptions.[12]

The second lesson is due to the particular pragmatist account of truth that is entailed by the scientific method. This view is not a naïve correspondence account by which we should somehow be able to compare our beliefs (or their contents) with "reality". Neither is this account a mere explication of how, in practical terms, an in-built fit between our beliefs and the world can be achieved or recognized. Rather, what it practically speaking *means* for our opinions to accord with an independent reality is itself to be worked out in a concrete fashion. Even Peirce's suggestion that truths are those beliefs that would withstand doubt were science to be pursued indefinitely implies nothing by way of the methods that are to be deployed to attain

[11] A common objection to this picture maintains that the central notion of an end of inquiry – or what it would mean for an opinion to withstand all future inquiry – is impossible to grasp, and that this will render the pragmatist view hopelessly murky. This objection, however, rests on a confusion between the abstract and the particular. It is not inherently difficult to abstractly conceive of what it would mean for an opinion to be sustained even at the end of inquiry pushed indefinitely. On the other hand, however, there is no way for us to tell that we have, on any particular question, reached the end. But this is only to be expected: the scientific method implies a thoroughgoingly fallibilist attitude towards any hypothesis.

[12] This view differs from that of some contemporary Peircean pragmatists – most notably Cheryl Misak (2000) and Robert B. Talisse (2007, 2010) – who have maintained that the notion of truth embedded in Peirce's scientific method is inevitable due to the nature of belief itself: belief can be genuinely fixed by scientific means only. While not relying in their on an analysis of the concept of truth (which they rightly note the pragmatists did not intend to supply), these defences of the scientific method rely on an analysis of the concept of *belief*. But Peirce nowhere suggests that the opinions fixed by methods other than the scientific one are less than genuine beliefs (cf. Rydenfelt 2011b).

this goal. The particular methods and of science – norms and desiderata for inquiry and theories – are themselves open to revision.

6 Non-representationalist Realism

Finally we have reached the point where the pragmatist conception of truth, including the hypothetical realism just developed, can be woven together with the non-representationalist point of view with which we started out. In its global form, expressivism maintains that claims made in any discourse or domain do not represent or describe in a straightforward fashion. It loosens the grip of the representationalist picture by which our thoughts and claims are automatically "about" something, or intended to "fit" truthmakers that are there in the world. As I have argued, the pragmatist, too, does not maintain that our claims – or thoughts, beliefs and the like – are automatically "about" some independent or external reality.

The pragmatist *does* however hold that our practical stances – expressible by way of making claims – *may* be made to accord with such a reality. Again, this does not revert the pragmatist back to the representationalist picture. The scientific method does not imply that we can as if compare our stances with what they are about, or represent. Rather, science is conceived of as the project of attempting to uncover ways to make our opinion accord with reality. The features that suggest that a particular hypothesis accords with an independent reality is up to scientific *practice*. In a sense, as science progresses, we *learn* what our beliefs, thoughts and claims are "about". Pragmatism thus offers us a version of realism which goes along with non-representationalism. As we have seen, the pragmatist notion of realism does not contradict any of the basic tenets of the contemporary non-representationalist; it merely *complements* them.

In fact, the non-representationalist position as developed by Price already includes a conceptual space for the kind of realism advanced by the classical pragmatist. Price has pushed a distinction between two notions (or nodes) of representation, which he suggests should replace the standard picture. Price's *i-representation* (where "i" stands for "internal" or "inferential") covers the sort or answerability that comes with the expression of a commitment or stance inside a discourse. It is due to this sort of representation that for those involved in that discourse, it appears as if they were talking about the way things *are*). I-representation contrasts with another node of representation, *e-representation* (where "e" stands for "external" or "environment"). This type of representation is involved when something – say, a device for measurement of some sort – is intended to react to environmental conditions.

In Price's view, running these notions together, or thinking of one as the primary node of representation, has resulted in the problems we face when trying to discover the facts "out there" that our claims and thoughts are supposed to match in the usual representationalist picture. While any discourse where notions such as "truth" or

"facts" are invoked is i-representational, it is only in genuinely e-representational discourses that the purpose of claims made is to react to, or covary, with things in our environment.

Following Price's terminology, the scientific method could now be understood as turning a discourse into an e-representational one: it attempts to make the claims made in that discourse answerable to an independent (or "external") reality. Importantly, there is no principled barrier to what kind of stances, claims, or discourses can be brought under the scientific fold. Price, on his own part, appears to maintain that e-representation covers much of claims made in contemporary natural science, while other domains, such as those of moral and modal claims, are i-representational only. However, this perspective still shows the remnants of old-fashioned representationalism. It threatens to make e-representation the direct conceptual offspring, a sort of a residue, of the "representationalist" view that Price rejects, so that the lines between what is i-representational and what is (also) e-representational will inevitably fall in the place where the borders between local expressivism and descriptivism already used to lie. This picture relies on a conception of the sort of facts that we may encounter derived from our contemporary understanding of natural science and the objects of its investigations, taking for granted that scientific realism must be realism of science as we now conceive it and delimiting the scientific enterprise to its current image.

Instead, my suggestion here is that the moral to be drawn from global non-representationalism is even more radical. Peircean *hypothetical* realism does not entail a commitment to any particular ontological picture: it is not a realism about the results of science, past, contemporary or future, but about an independent reality which our claims may accord with. Assuming such realism, the scientific enterprise can be extended to domains that contemporary scientific practice leaves untouched. In a discourse turned e-representational – as opposed to i-representational "only" – the opinions expressed are not merely open to criticism from a point of view with transcends the speaker's subjective opinion, but more specifically answerable to the standard of an independent reality. (This *norm* of answerability to such a standard external to the discourse is itself unavoidably internal to the discourse, but this does not make the *standard* any less external.) Obviously, the specific nature of the *reality* in question depends on the discourse at hand, and giving a plausible account of it will require scientific discovery and (philosophical) conceptual work.

7 Normative Realism and Normative Science

It is this insight that finally brings us to the idea of *normative realism*, which will serve as a case in point. Forms of normative realism (such as *moral* realism) have standardly been conceived of as the combination of two theses. The first is the *cognitivist* semantic thesis: it maintains that normative (or moral) judgments are fact-stating, or describe ways things are. The second thesis is ontological: it holds that there are things such as described by (some) normative judgments. As a

third component, many moral realists have insisted that these (moral) facts must be independent of what we think, believe, desire and so on.

The cognitivist thesis faces two major difficulties. The first is the problem of accounting for the facts our normative claims are supposedly "about". After Moore, the cognitivists have either retreated to forms of non-naturalism about normative "facts" (Shafer-Landau 2003), or attempted to give viable (naturalist) accounts of the conceptual content or the reference of normative predicates (e.g. Smith 1995; Boyd 1988). But the former alternative implies the existence of strange non-natural facts, which fit uneasily into the scientific and philosophically naturalist world-view; and the latter accounts have not arrived at any commonly accepted results. The second difficulty is that normative claims and thoughts appear to play a different role in our agency than that of non-normative claims and thoughts. By contrast to "descriptive" claims, normative claims tell of the outcomes we aim at and the sort of actions we are prepared to promote or avert, praise or reprimand – the *ends* or *goals* of our actions.[13] The cognitivist position as commonly conceived has, if anything, fuelled scepticism about the normative: it appears we have no plausible account of what sort of facts normative claims are "about" in the first place, or at the very least much reason to doubt the existence of such facts.

The non-cognitivist alternative sets out to deal with both these problems with a simple and elegant response. It maintains that the cognitivist project is futile as normative claims do not describe the world; instead, such claims express such functional states that play the relevant practical role of setting the ends or purposes of action. Perhaps the most central difficulty of this view is its unsettling implication that there is nothing to back our normative views beyond the preferences we merely happen to have – a form of *relativism* that this position results in. To be sure, the non-cognitivist position is not a form of what we could call *conceptual* relativism (cf. Horgan and Timmons 2006). Quite the contrary, it contests the view that normative claims or terms refer to the conative states of those who make such claims.[14] Neither does the non-cognitivist maintain that any normative or moral view is as good as any other: this would amount to a normative (or moral) stance of its own right, and arguably a very strange one at that (cf. Blackburn 1998, p. 296). But as we already saw, proponents of expressivism have drawn attention to the demand of intersubjective agreement in many of our discourses (Price 1998, 2003; cf. Gibbard 2003, ch. 4); and debates over normative issues count among them: differences in moral opinion certainly invite disagreement. If our normative and

[13]This idea is often cast in terms of moral motivation: normative claims are thought to have a distinct, conceptual connection to what we are motivated to do. This fact has caused problems for the cognitivist view when coupled with the so-called Humean theory of motivation, which maintains that beliefs as such are not sufficient for motivation, but require the presence of other mental states or dispositions, commonly called desires.

[14]Indeed, expressivists have long argued that their view does not attempt to give conceptual content for normative vocabulary in terms of the attitudes of the speaker, his group, or the like. Expressivist non-cognitivism is thus crucially different from views such as (moral) speaker subjectivism, which maintains (for example) that to call an act wrong is to say that one disapproves of that act.

moral preferences are simply the products of our personal development as well as that of our societies, and there is nothing beyond them to settle our common opinion, what are our hopes of attaining a lasting consensus over normative affairs?

Non-representationalist realism – the pragmatist view here developed – can avoid both the sceptical and relativist concerns of the customary alternatives. Equipped with the representationalist picture, the traditional cognitivist has been looking for a match between normative claims (or their conceptual contents) and "facts", leading to sceptical results. Abandoning the representationalist view, non-cognitivism has been taken to eschew all grounds for a lasting consensus, raising the worry of relativism. The pragmatist, however, conceives of the answerability of opinion to reality more broadly: it does not require of our opinions to *represent* reality in order to be *answerable* to it. It is by abandoning the representationalist assumptions while reconceptualizing realism that the pragmatist view developed here can bring normative and non-normative claims under the same fold. In the global non-representationalist view, neither are at bottom any more (or less) "cognitive": the difference between these claims pertains to their different functions in discourse and action rather than in their "representational" capacities. By adopting the scientific method, *both* kinds of opinions may be settled in a manner that is sensitive to (an independent) reality.

Even when relieved of the burden of giving an account of what our normative claims and thoughts "represent", the pragmatist will still need to supply a view of the sort of *reality* that our normative opinions can be made to accord with, and how that reality may affect us as inquirers – that is, an account of the form that hypothetical realism could take in normative matters. Fortunately the pragmatist has at hand at least the beginnings of such an account Peirce's later views, which have been further developed and elaborated by T. L. Short (2007). Peirce recognized that teleology had been reintroduced to modern science in that some forms of statistical explanation are not reducible to mechanistic causation. As an extension of this naturalistic view of final causation, he suggested that certain ideas (or ideals) themselves may have the tendency of becoming more powerful by gaining more ground. Our normative opinions are to be settled in accordance with such tendencies, independent of but affecting our particular inclinations and desires. As Short construes Peirce's later semiotic view, these ideals can affect us through experience by eliciting feelings of approval and disapproval, satisfaction and dissatisfaction. Experience may correct our feelings, and eventually force convergence among inquirers. These notions form the basis of Peircean normative science.[15]

This picture is admittedly sketchy, and will doubtless seem outlandish to many. But at bottom it only requires openness to the hypothesis that normative and moral ideals can, analogously to our non-normative opinion, be guided by an independent reality. The historical development and spreading of certain ideals – say, those concerning basic human rights and liberties, freedom of opinion, and the way we

[15]Similarly, the Peircean pragmatist may argue that the scientific method itself as well as the particular norms of science are due to the compelling force of experience (cf. Rydenfelt 2011a).

are to settle disagreements over matters of (non-normative) opinion – may be taken as evidence for the possibility of reaching a consensus, slowly and over time, about such issues. Not much more can be said about the nature of the reality that normative science pertains to, as otherwise too much in particular will be said about how ideals are to be settled and what kind of ideals would prevail – too much will be said about the *results* of such a science, rather than its foundations merely.

8 Conclusion

Contemporary expressivism contests the traditional idea that our thoughts and claims attempt to "fit" something in the world. This approach, when extended in Huw Price's fashion, results in a global non-representationalism. I have argued that the views of the classical pragmatists, especially Charles S. Peirce's account of inquiry and truth, are amenable to an expressivist and non-representationalist interpretation. The prevalent assumption that this cannot be the case is due to the received wisdom that realism – which the pragmatists advanced in different forms – entails representationalism. But for Peirce, the scientific method and its commitment to a hypothetical realism is not derived from conceptual considerations, or a robust notion of representation; instead, it is the outcome of a substantial development of ways of fixing opinion. As such it is fully compatible with the non-representationalist view: it is a realism without representationalism. One of the advantages of this novel, combined pragmatist perspective is that it enables us to radically reconceptualize other brands of realism, such as *normative* realism. Once we have adopted the global expressivist perspective, there is no principled, "representational" difference between normative and non-normative (or "descriptive") claims or opinions, and the pragmatist may argue that both kinds of opinion are to be fixed by the same – scientific – means.[16]

References

Ayer, A.J. 1952. *Language, truth and logic*. New York: Dover Publications.
Blackburn, S. 1988. Attitudes and contents. *Ethics* 98: 501–517.
Blackburn, S. 1998. *Ruling passions*. Oxford: Clarendon.
Boyd, R. 1988. How to be a moral realist. In *Essays on moral realism*, ed. G. Sayre-McCord, 181–228. Ithaca: Cornell University Press.
Brandom, R.B. 1994. *Making it explicit: Reasoning, representing, and discursive commitment*. Cambridge, MA: Harvard University Press.
Brandom, R.B. 2000. *Articulating reasons: An introduction to inferentialism*. Cambridge, MA: Harvard University Press.

[16]I'm grateful to Jonathan Knowles, Sami Pihlström, Huw Price, T.L. Short and Kenneth R. Westphal for comments on earlier versions of this paper.

Devitt, M. 1991. *Realism and truth*, 2nd revised ed. Oxford: Basil Blackwell.
Gibbard, A. 1990. *Wise choices, apt feelings*. Cambridge, MA: Harvard University Press.
Gibbard, A. 2003. *Thinking how to live*. Cambridge, MA: Harvard University Press.
Haack, S. 1998. *Manifesto of a passionate moderate*. Chicago: University of Chicago Press.
Horgan, T., and M. Timmons. 2006. Expressivism, Yes! Relativism, No! In *Oxford studies in metaethics*, vol. 1, ed. R. Shafer-Landau. Oxford: Oxford University Press.
James, W. 1897. *The will to believe and other essays in popular philosophy*. New York: Longmans Green.
James, W. 1907. *Pragmatism*. Cambridge: Harvard University Press. 1975.
Miller, A. 2002. Realism. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta, Spring 2012 ed. http://plato.stanford.edu/archives/spr2012/entries/realism/.
Misak, C. 2000. *Truth, politics, morality. Pragmatism and deliberation*. London: Routledge.
Moore, G.E. 1903. *Principia Ethica*. New York: Cambridge University Press.
Peirce, C.S. 1877. The fixation of belief. In *The essential Peirce*, vol. I, ed. N. Houser and C. Kloesel. Bloomington: Indiana University Press, 1992.
Peirce, C.S. 1878. How to make our ideas clear. In *The essential Peirce*, vol. 1, ed. N. Houser and C. Kloesel. Bloomington: Indiana University Press, 1992.
Peirce, C.S. 1903. The three normative sciences. In *The essential Peirce,* ed. the Peirce Edition Project, vol. II. Bloomington: Indiana University Press, 1998.
Pihlström, S. 2007. Brandom on pragmatism. *Cognitio* 8(2): 265–287.
Price, H. 1998. Three norms of assertibility, or how the MOA became extinct. *Philosophical Perspectives* 12: 41–54.
Price, H. 2003. Truth as convenient friction. *Journal of Philosophy* 100: 167–190.
Price, H. 2011a. Expressivism for two voices. In *Pragmatism, science and naturalism*, ed. J. Knowles and H. Rydenfelt, 87–113. Berlin: Peter Lang.
Price, H. 2011b. *Naturalism without mirrors*. Oxford: Oxford University Press.
Putnam, H. 1981. *Reason, truth, and history*. Cambridge: Cambridge University Press.
Rorty, R. 1982. *Consequences of pragmatism*. Minneapolis: University of Minnesota Press.
Rydenfelt, H. 2009a. The meaning of pragmatism. James on the practical consequences of belief. *Cognitio* 10(1): 81–90.
Rydenfelt, H. 2009b. Pragmatism and the aims of inquiry. In *Pragmatist perspectives*, ed. S. Pihlström and H. Rydenfelt, 41–52. Helsinki: Acta Philosophica Fennica.
Rydenfelt, H. 2011a. Naturalism and normative science. In *Pragmatism, science and naturalism*, ed. J. Knowles and H. Rydenfelt, 115–138. Berlin: Peter Lang.
Rydenfelt, H. 2011b. Epistemic norms and democracy. *Metaphilosophy* 42(5): 572–588.
Rydenfelt, H. forthcoming. Constructivist problems, realist solutions. In *Persuasion and compulsion in democracy*, ed. J. Kegley and K. Skowronski. Lanham: Lexington Books.
Schroeder, M. 2008. *Being for*. Oxford: Oxford University Press.
Shafer-Landau, R. 2003. *Moral realism: A defence*. Oxford: Oxford University Press.
Short, T.L. 2007. *Peirce's theory of signs*. Cambridge: Cambridge University Press.
Smith, M.J. 1995. *The moral problem*. Oxford: Blackwell.
Stevenson, C. 1944. *Ethics and language*. New Haven/London: Yale University Press.
Talisse, R.B. 2007. *A pragmatist philosophy of democracy*. London: Routledge.
Talisse, R.B. 2010. Peirce and pragmatist democratic theory. In *Ideas in action: Proceedings of the applying Peirce conference*, ed. M. Bergman, S. Paavola, A. Pietarinen, and H. Rydenfelt, 105–116. Helsinki: Nordic Pragmatism Network.
van Roojen, M. 2004. Moral cognitivism vs. non-cognitivism. In *The Stanford encyclopedia of philosophy*, ed. E.N. Zalta, Spring 2011 ed. http://plato.stanford.edu/archives/spr2011/entries/moral-cognitivism/.

The Continuing Relevance of Nineteenth-Century Philosophy of Psychology: Brentano and the Autonomy of Psychological Methods

Uljana Feest

1 Introduction

One prominent theme in the philosophy of psychology concerns the status of psychology vis-à-vis neighboring disciplines, such as neurophysiology. While much of the literature revolves around the question of whether psychological *explanations* are reducible to neuroscientific explanations, this issue can be distinguished from another question, i.e., whether psychological *research* can proceed in a manner that is independent of research conducted in neuroscience. In a rough-and-ready fashion, we may say that the former question is about a "static" state of affairs, i.e., what is the relationship between the explanations provided by psychology and neuroscience, whereas the latter address a *dynamic process*, i.e., the process of research. We may refer to the former issue as concerning "explanatory autonomy," whereas the latter concerns "methodological autonomy".[1]

I have maintained in the past that these two issues can be separated, and I argued against methodological autonomy (Feest 2003). The basic gist of this argument was to say that in the process of forming their taxonomies, "higher" and "lower" level sciences mutually inform one another. This argument is similar to some in the recent literature in the philosophy of psychology and neuroscience (Bechtel and Richardson 1993; Craver 2007), which have also pointed out that there is often a dynamic interaction between psychological and neuroscientific research. My claim (and that of others) that psychology is not (and cannot be) methodologically

[1] See (Feest 2003). Retrospectively, it seems to me that a more accurate term for what I had in mind would have been "procedural autonomy".

U. Feest (✉)
Leibniz Universität Hannover, Institut für Philosophie, Im Moore 21, 30167 Hannover, Germany
e-mail: ufeest@yahoo.com

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_48,

autonomous can also be cast in a more positive fashion, as saying that psychology is (or should be) methodologically (or procedurally) *integrated* with other sciences.

Shifting the emphasis away from a negative thesis (against autonomy) and towards a positive thesis (for integration), brings to the fore the question of *what exactly is supposed to be integrated*. In the recent literature, it is typically assumed that the relata of psycho-neural integration are *explanations* (see for example Piccinini and Craver 2011). By contrast, and in keeping with my interest in the investigative procedure as such, I want to raise the question of whether psychology and its neighboring disciplines employ unique and distinctive *methodologies*. This question highlights the fact that we need to distinguish between *two* issues with regard to *methodological autonomy vs. integration*: one concerns the question of whether *the process of knowledge generation* in psychology can be autonomous from that of neuroscience. The other concerns the question of whether psychology contributes its own *distinctive methodology* to this process. While continuing to reject methodological autonomy in the former sense, I will tentatively adopt a version of an argument in favor of methodological autonomy in the latter sense. In this vein I will suggest that the autonomy provided by the existence of distinctive methodologies would not be an impediment, but a prerequisite for genuine integration, because it would contribute to the production of the very findings that need to be integrated in neuropsychological research.

In this paper I will focus on a class of methods that in the nineteenth century was by many viewed as distinctly psychological, i.e., methods that make use of first-person data. One person who explicitly linked his discussion of such methods to the issue of the autonomy of psychology was Franz Brentano (1838–1917). Brentano's views, I will argue here, are still well worth considering, in particular as he explicitly related practical questions about psychological research to underlying metaphysical commitments. I will begin with a brief overview of Brentano's vision of psychology (Sect. 2) and then present evidence for my claim that Brentano was concerned with issues of methodological autonomy (Sect. 3). I will outline his views about the method and subject matter of psychology as underwriting a specific response to this concern (Sect. 4). I will then show that we can attribute to Brentano the idea that methodological integration presupposes methodological distinctiveness, and the insight that questions about the choice of methods are deeply tied to metaphysical considerations about the subject matter of psychological investigations (Sect. 5). The paper will conclude by arguing for the continuing relevance of these considerations to more recent philosophy of psychology (Sect. 6).

2 Brentano's Methodological Views in and Out of Context

In his 1874 book *Psychology from an Empirical Standpoint* Franz Brentano distinguished between two kinds of psychology, *descriptive* and *genetic* psychology. The former had the task of providing phenomenological descriptions of mental phenomena. The latter had the task of explaining them. However, as he emphasized

in many places, for him the focus of genuinely *psychological* endeavours was on the *description*, not the explanation, of mental phenomena. The reason for this was that according to him the explanations provided by a genetic psychology would inevitably appeal to unconscious and/or physiological states or processes, whereas a psychological explanation could only make reference to mental phenomena, which he defined as *conscious*. Thus, Brentano's rejection of the possibility of "purely" psychological explanation rested on a very specific understanding of what such explanations, if they were to exist, would look like, i.e., they would appeal to laws connecting conscious mental states to one another. Since he did not think that many such laws were likely to be found, he focused his efforts on thinking about how to provide adequate descriptions of what we might refer to as the "explanandum phenomena" of psychology, i.e., conscious mental states. The explanations provided by genetic psychology would then appeal to a hybrid of mental and physiological states and processes.

Given Brentano's understanding of "mental phenomena" as conscious mental states, and given his understanding of descriptive psychology as aiming to provide phenomenological descriptions of such states, it is not surprising that he thought that psychology had to devise *specific methods* that would allow it to pursue its task. This method contained as an ineliminable component something he called "inner perception," i.e., a form of first-person access to the objects in question. Thus, Brentano thought that inner perception was a necessary (though, as we will see, not sufficient) condition of the very possibility of an empirical psychology. In addition he held that by employing inner perception, psychology differed in principle from other empirical sciences, which involved what he called "outer perception".

Returning to our earlier distinction between the issues of explanatory and methodological autonomy of psychology, it is tempting to say that Brentano rejected explanatory autonomy but endorsed methodological autonomy. This is a somewhat fair characterization, though I will suggest that his thesis is more accurately described as one about the *methodological distinctiveness* of psychology since he did not believe that psychological research could take place in complete separation from neurophysiology, but merely that it employed an irreducible and unique method, which could be applied in concert with other methods.[2] In other words, I argue that he viewed the existence of distinctive methods as compatible with the possibility of *methodological integration*. This may seem like a reasonable position to take, but it bears stressing that Brentano had very specific arguments for it, and that they occurred in the context of very specific debates. So, how should we evaluate Brentano's position from a contemporary perspective?

In the following, I will begin with a reconstruction of the views in question, which contextualizes them vis-à-vis Brentano's systematic philosophy of mind as well as the debates he was engaging in at the time. I do this because I believe that we are in danger of missing important features of a historical writer's position if we don't understand the specific intellectual contexts in which he

[2]I will explain and provide evidence for this claim in Sect. 5 below.

formulated them. However, this raises questions about the transferability of this position to contemporary debates. Differently put, if we need to contextualize a philosophical argument in order to understand it better, does this mean that the insights articulated in that argument are not applicable outside the specifics of their original formulation? While this issue will not be at the foreground of this article, we will touch on it towards the end. In general terms, I argue that questions about the transferability of specific aspects of a past system of ideas can only be asked on a case-by-case basis, and the answer will depend on the specific questions we attempt to attack by appeal to the ideas of a past thinker. In this vein, it is not the aim of this article to advocate a whole-sale revival of Brentano's philosophy of psychology. Rather, I will show that Brentano's approach offers a unique and original perspective to present-day discussions of methodological autonomy/integration and of the status of first-person data to phenomenological descriptions.

3 The Boundaries of Psychology, and Why They Mattered to Brentano

Brentano's work about the distinctiveness of psychological methods had two targets: (1) he wanted to provide the foundations of a methodological unification *within* psychology, and (2) he wanted to provide a foundation for psychology *as distinct and/or autonomous from* neighboring disciplines. In this vein, I find it helpful to think of his reflections about methodology as "boundary work" (Gieryn 1983).[3] Let me be clear that while Brentano took an explicit interest in the question of boundaries between psychology and other sciences, he did not think that such boundaries were cast in stone. In this vein, he talked about the issue of "Grenzziehung" (boundary drawing) and even "Grenzstreitigkeiten" (border quarrels) between psychology and physiology (Brentano 1874/1973, p. 7), but acknowledged that every division of scientific fields, no matter how good, will be somewhat artificial (*id.*, p. 8).[4] Still, he claimed, even in border-disciplines, such as psychophysiology and psychophysics, it is possible to say, for specific questions, whether they are to be approached by the methods of psychology or physiology. The reasoning for this was that for Brentano the unique task of describing mental phenomena by means of inner perception delineated psychology quite clearly from other tasks.

[3]While Giery's initial analysis was primarily aimed at the ways in which boundaries between science and pseudoscience are established, he later on argued that "[t]he utility of boundary-work is not limited to demarcations of science from non-science. The same rhetorical style is no doubt useful for ideological demarcations of disciplines, specialties or theoretical orientations within science" (Gieryn 1983, p. 792).

[4]While I haven't researched this very thoroughly, I gather that the metaphor of borders as delineating psychology from physiology and philosophy was not uncommon at the time. In this vein, Bordogna (2008) provides evidence that both William James and Wilhelm Wundt used this terminology in the 1860s and 1870s.

3.1 Defending the Border to Physiology

In Chap. 3 of *Psychology from an Empirical Standpoint,* Brentano argued that even though the ultimate aim of psychology was to determine the laws of succession of mental phenomena, it was unlikely that exceptionless laws could be found. He explained that this was due to the fact that psychological phenomena depend on a great number of physiological conditions of which we have only incomplete knowledge. However, Brentano hastened to add that by this he did not mean to suggest that the laws of mental succession could be derived from those of physiology, even if we knew those. Clearly, mental phenomena are dependent on physiological phenomena, but this does not mean that psychological categories could be derived from physiology, an assumption he attributed to Gall, Comte, and Horwicz (Brentano 1874/1973, p. 60). Even if we acknowledge that physiology may have a role to play in psychology, Brentano argued, the genuinely psychological method was going to have an ineliminable component. Two authors were targeted in particular:

The first one was the German philosopher Adolf Horwizc (1831–1894), who had recently laid out his views in the first of two volumes of a book, entitled *Psychologische Analysen auf physiologischen Grundlagen* (1872a) and an article, "Methodologie der Seelenlehre" (1872b). There he had argued that while an introspective analysis of consciousness might well function as a heuristic device, the real work of delineating mental taxonomies and laws was ultimately going to be done by physiology. Brentano's critique of this was twofold. First, he questioned that the vision Horwizc had of the relationship between "lower" and "higher" level sciences ever held true, even outside of psychology. For example, while it is surely uncontroversial that inorganic chemistry and physics might be of help to physiology, nobody would expect to derive any knowledge of physiological structures from these other sciences. Inorganic processes are necessary for organic ones, Brentano argued, but the nature of the organic processes will have to be investigated in their own right. Second, according to Brentano, even if it *were* true for the relationship between inorganic chemistry and physiology, the analogy would break down when it came to physiology and psychology, for the simple reason that physiology dealt with external phenomena, psychology with internal phenomena (Brentano 1874/1973, p. 64). This distinction between inner and outer phenomena is of course highly problematic as was pointed out early on by Brentano's student Edmund Husserl (see also Feest 2012b). Nonetheless, I will argue below, Brentano pointed to an important feature of the psychological investigation of subjective experience, i.e., its reliance on *first-person methods*. The question of whether such methods are adequately described as having "inner phenomena" as their objects should not detract us from the systematic status of first-person data, including the question of how they are generated and what are the limits of their utility to scientific investigations.

Brentano's second target was the British psychiatrist Henry Maudsley (1835–1913), who, in his 1866 *Physiology and Pathology of the Mind,* had argued that not all mental phenomena are conscious, and that therefore physiological facts are required for any attempt to describe (let alone explain) the totality

of facts about mental phenomena (Maudsley 1867). In addition, even if we succeeded in formulating laws of the succession of conscious states, Maudsley had argued, these would still require physiological explanations. In response to these arguments, Brentano (unsurprisingly) questioned Maudsley's claim that not all mental phenomena are conscious. This critique followed directly from his conception of mental phenomena as conscious. (Section 4 will give a more detailed treatment of Brentano's arguments for this conception.) Furthermore, while Brentano believed it to be unlikely that there are many laws that can be established between successive conscious mental states, he granted Maudsley that if such laws were to be found they might need to be explained by appeal to physiological processes. However, he argued that in this respect psychological laws are no different than – say – the empirical laws of mechanics, and that a genuinely psychological method was still required to establish the empirical laws to begin with.

3.2 Defining Psychology Against Other Visions of Psychology

In addition to defending the boundary between psychology and physiology, Brentano also engaged in *internal fortification* by unification. In this vein, he stated in the preface to his *Psychology*, that he sought "to establish a single unified science of psychology in place of the many psychologies we now have" (Brentano 1874/1973, p. xvi). As we just saw, Brentano drew the boundary between psychology and physiology by (a) limiting the task of psychology to that of giving empirical taxonomies and laws, and (b) arguing that these tasks could only be accomplished by means of a method based on inner perception, which is unique to psychology. In defining the scope and methods of psychology in this way, however, Brentano positioned himself vis-à-vis other approaches ("psychologies") at the time, which took the task of psychology to be not only that of providing descriptions, but also *explanations* of conscious mental phenomena, and/or which debated different methods of gaining epistemic access to conscious mental states.

There is a sense in which it was widely held in the nineteenth century that consciousness was the proper subject matter of psychology. However, opinions differed over (a) whether conscious mental phenomena were to be explained or merely described by psychology, and (b) what were legitimate scientific methods for gaining access to those phenomena. To explain what I mean by these points, let me use the example of Wilhelm Wundt, whose *Grundzüge der physiologischen Psychologie* appeared in the same year as Brentano's *Psychologie von einem empirischen Standpunkt* (Wundt 1874/1973–1974). For Wundt, sensory consciousness was an important object of psychological research and he (like many others at the time) held that phenomenally conscious mental states could be explained by appeal to basic or immediate "elements" of consciousness and some mental processes (e.g., association, or – in his case – apperception), which constructed conscious experiences out of the elements. As indicated above, Brentano did not

find this type of explanatory approach promising. One major reason for this was that we are typically not conscious of the supposed elements in question, and hence it was not clear that these explanations were really appealing to mental phenomena, and that they were psychological explanations at all.

The question, then, was what kind of evidence someone like Wundt could produce in support of his research program. As Wundt was going to spell out in a later publication (Wundt 1888), he realized that under ordinary circumstances we do not have introspective access to the basic elements of sensations posited by his explanatory approach (i.e., he knew that the units of our ordinary conscious experience are typically "larger" and more complex). However, he believed that it was possible to create highly controlled experimental conditions under which the supposed basic elements could be made accessible and introspectively reported (Hatfield 2005). As will be explained more fully in the following section, Brentano did not believe concurrent introspection to be a viable method, so this would have been a clear point of disagreement between them. While I cannot provide a detailed discussion of this here, I claim that the disagreements at hand were representative of a divide that pervaded late nineteenth-century German psychology more generally, thus underwriting my thesis that the (broadly) Wundtian and elementist approach to psychology was one of Brentano's targets when aiming "to establish a single unified science of psychology".

4 Method and Subject Matter: The Place of Inner Perception

In the previous Section I argued that Brentano's writings about the methods of psychology had specific contexts, and that they have to be understood as *boundary work* that simultaneously aimed to unify psychology from within and demarcate it from other pursuits, such as physiology. Brentano did the latter by highlighting an aspect of psychological methodology that he deemed to be unique, namely the reliance on inner perception: "[I]nner perception of our own mental phenomenon … is the primary source of the experiences essential to psychological investigations" (Brentano 1874/1973, p. 34). However, up to now I have not explained what he meant by "inner perception", how it was to figure in his methodology, and how this methodology was anchored in Brentano's theory of mind.

4.1 Inner Perception, Inner Observation, and Retrospective Inner Observation

It is important to understand that Brentano did not think that inner perception, taken by itself, constituted a *method*. Instead, he only held that it would contribute the "raw material" for a psychological method. Moreover, he explicitly distinguished *inner perception* from *introspection*: "Note … that we said that inner perception [Wahrnehmung] and not introspection, i.e., inner observation [Beobachtung]

constitutes this primary and essential source of psychology" (*id.*, p. 29). Indeed, he thought that whereas inner perception constituted the "primary source" of psychology, introspection (understood as *concurrent* introspection of the material present in inner perception) was highly problematic as a method since it would distort what was being observed: the perceptual state in question. But if concurrent introspection was inadmissible as a psychological method, what would an adequate psychological method look like that contained inner perception as an essential component?

Brentano's answer was that we can have access to the mental states revealed by inner perception by *retrospective* inner observation. In other words, he thought that they could be observed *in memory*. This followed from Brentano's conceptions of *inner perception* and *scientific observation*, respectively. According to Brentano, the difference between inner perception and any kind of scientific observation was that inner perception was immediate and infallible, whereas scientific observations are detached and fallible. This disqualified inner perception from ever counting as inner observation since it was by definition infallible. Having argued that concurrent inner observations were not suitable, this left retrospective observations of inner perceptions as the method of choice. In practical terms, therefore, the recommendation was to conduct experiments and question subjects about their experiences after the experiment. Brentano realized of course that memory can distort and misrepresent what the experience was really like: "As everyone knows, memory is, to a great extent, subject to illusion, while inner perception is infallible and does not admit of doubt" (*id.*, p. 35). Hence, he saw a trade-off between the immediacy and infallibility of inner perception and the observability of the mental phenomena in memory, but argued that it is better to have fallible observations than no observations at all, noting that fallibility is a feature of scientific observations in general.

4.2 Mental States as Conscious Phenomena and Only as Conscious Phenomena

Brentano's vision of the necessity of inner perception depended crucially on the assumption that this method had full access to *all* mental phenomena. Justifying this assumption required spelling out his notion of *mental phenomenon*. Brentano did so in Book II of his *Psychology from an Empirical Standpoint*, entitled "Mental Phenomena in General". It is in this context that Brentano famously introduced the notion of the *intentionality of mental phenomena*, defining a mental phenomenon as "consciousness of an object". He went on, however, to acknowledge that this formulation did not in and of itself guarantee that all mental states were accessible to inner perception (and ultimately to retrospective inner observation), since it was after all possible to be conscious of an object but not to have conscious access to this intentional mental state:

> We have seen that no mental phenomenon exists which is not … consciousness of an object. However, another question arises, namely, whether there are any mental phenomena which are not objects of consciousness. All mental phenomena are states of consciousness; but are all mental phenomena conscious, or might there also be unconscious mental acts [?] (*id.*, p. 102)

As he put it further down on the same page: Might there be an "unconscious conscious"? The issue at stake was whether there could be intentional states (consciousness of an object), which were in principle inaccessible to what is today sometimes referred to as "higher-order consciousness".[5] Brentano rejected the very distinction between higher- and lower-order consciousness, and as we will see in a moment, his reasoning for this underwrote his views about *inner perception vs. inner observation*. At the same time, however, he did not reject questions about an "unconscious conscious" out of hand, remarking that different versions of it had been around for some time (Aquinus, Leibniz, Kant, Mill, Herbart, Helmholtz and others). In trying to refute these, Brentano acknowledged that since the hypothesis he was attacking was that there are (at least) some unconscious mental phenomena, it could not be refuted simply by showing that there is no introspective experiential evidence for it. Still, Brentano thought that the very notion of unconsciously consciousness phenomena was misguided, and he put forward several arguments in support of this conviction. We will only briefly outline them here, in order to highlight the extent to which his methodological reflections on this point were entangled with his philosophy of mind.

Brentano's first strategy was to formulate, and then refute, several possible reasons for the thesis of the existence of unconscious mental processes. I will only gesture at the general strategy adopted here. Essentially, Brentano argued that before we appeal to unconscious mental phenomena as explanatory of conscious mental phenomena, we need to establish (a) that the phenomenon that is explained by appeal to unconscious phenomena in fact exists, and (b) that there is no better explanation for it. For example, when someone says that they have suddenly become conscious of a feeling of love that they have unconsciously harboured for some time, this might be considered a phenomenon that calls for a notion of unconscious mental acts. However, Brentano disagreed with this description and stated "The truth is that we were conscious of each individual act when we were performing it, but that we did not reflect upon it in a way that allowed us to recognize the similarity between the mental phenomenon in question and those which are commonly called by its name" (*id.*, p. 115). Hence, Brentano introduced an important conceptual clarification by arguing that if a given feeling can in principle be brought into consciousness it was conscious all along. For Brentano, a mental state was genuinely unconscious only if we could not make it conscious by directing one's attention to it. But if it was genuinely unconscious in this sense, he maintained, it had to be regarded as a physiological rather than a mental phenomenon.

[5] I am thinking here of the positions of philosophers such as David Armstrong; William Lycan and David Rosenthal.

Brentano's second strategy was to cast doubt on the existence of an epistemic gap between mental phenomena and our perception of them. Essentially, he suggested that if it were the case that some conscious mental states were inaccessible, or only partially or incorrectly accessible to higher-order conscious mental states, this would imply that we can be wrong about our conscious mental states, and this would go against Brentano's dictum of the infallibility of inner perception. In turn, this dictum was explained in his third argument against the possibility of unconscious mental states. There he cast doubt on what might at first glance appear to be a reasonable construal of how inner perception works, i.e., that inner perception gives rise to a higher-order mental state that has as its intentional object a lower-level perception. As Brentano pointed out, this model has the paradoxical consequence that the object in question (for example, a sound) is presented twice. To this he answered that it just does not square with our experience, and that therefore the distinction between a presentation and a presentation of a presentation is at best analytically useful. For example, "[t]he presentation of the sound and the presentation of the presentation of the sound form a single mental phenomenon" (*id.*, p. 127).

Notice, however, that this argument was begging the question since Brentano had already acknowledged that appeal to experience will not satisfy those who believe that not all conscious mental states are accessible to experience. Nonetheless, his argument is useful in highlighting the extent to which his views about the methods of psychology were intertwined with his philosophy of mind: For him, inner *perceptions* were infallible, because they were intrinsically tied to their objects (hence no epistemic gap), whereas inner *observations* were conceptually impossible, because the mental state of the observer and the observed mental state are literally the same.

5 Brentano's Argument for Methodological Distinctiveness Did Not Rule Out Integration!

In the previous two Sections I laid out two contexts for Brentano's thesis of the methodological distinctiveness of psychology: (1) Existing debates at the time about the methods and boundaries of psychology and (2) the metaphysical underpinnings of Brentano's own specific contribution to those debates. As we saw, Brentano had a very specific vision of the subject matter of psychology (phenomenally conscious mental states). In turn, this specific vision also motivated his argument in favour of a *distinctly psychological method*, namely one that relied on first-person access to phenomenally conscious mental states. In addition, his philosophy of mind placed constraints on the form that this method could take: Both the infallibility of *inner perception* and the impossibility of *inner observation* were sophisticated implications of Brentano's theory of mind with its central thesis of the intentionality of mental states. As a result, his proposed *distinctive method of*

psychology (retrospective inner observation) was deeply and systematically tied to his philosophy of mind.

In this section, I want to substantiate my claim that Brentano's thesis of the *methodological distinctiveness* of psychology did not imply the *methodological autonomy* of psychology. I have argued that Brentano endorsed the view that (a) there is a method that distinguishes psychology from other sciences, and (b) this method is, ultimately, not eliminable from the study of the mind. However, he (c) did not think that this method was generally to be employed *in isolation from* other methods.

> [T]he experimental foundation of psychology [. . .] would always remain insufficient and unreliable, if this science were to confine itself to the inner perception of our own mental phenomena and to their observation in memory. This is not the case, however. In addition to direct perception of our own mental phenomena we have an indirect knowledge of the mental phenomena of others. The phenomena of inner life usually express themselves, so to speak, i.e., they cause externally perceivable changes. (*id.*, p. 37)

Brentano went on to explain that such "externally perceivable changes" included words and nonverbal forms of communications, as well as involuntary behaviors that indicate mental phenomena. Indeed, the very fact that verbal and nonverbal communication about our mental phenomena is possible, he argued, is evidence for the presupposition that the subjective experiences of individuals are not so different from one another, making it possible to treat individual retrospectively introspective data as representative of human mental phenomena more generally. Brentano emphasized, however, that "externally perceivable" data could never in of themselves be sufficient for a science of psychology. "It is not possible . . . that this external . . . observation of mental states could become a source of psychological knowledge, quite independently of inner 'subjective' observation" (*id.*, p. 40). If we add neurophysiological processes to Brentano's list of externally perceivable changes (and I see no reason why he would have objected to this), I suggest that we classify his position not as one that favours methodological *autonomy*, but rather as one that allows for methodological *integration*. In other words, he thought that a distinctly psychological method (retrospective inner observation) was an ineliminable component for the study of mental phenomena that would benefit from integrating methods of other fields as well.

One crucial insight I would like to draw from Brentano's considerations is that in order for there to be methodological integration, there have to be *distinct methods* in the first place. As such, it strikes me that Brentano's philosophy of psychology certainly has some systematic suggestions to offer that promise to be fruitful to questions and debates of more recent philosophy of science, especially regarding the nature of *integration*. Without going into much detail here, let me just briefly indicate what debates I have in mind: the past 20 or so years have seen some interest in the *disunity of science* both as an empirical fact and as a philosophical challenge (e.g., Galison and Stump 1996). In turn, this has given rise to attempts to give philosophical accounts of how the *plurality* of approaches and phenomena might be *integrated*. In many cases, such attempts have radically questioned traditional accounts of the unity of science, both by starting with detailed

investigations of scientific practice and by rejecting the traditional assumption that integration necessarily implies reduction (e.g., Mitchell 2003). More recently, there has also been significant interest in analyzing the nature of *interdisciplinarity* as well as the challenges of and inter- and multidisciplinary collaborations (see Andersen and Wagenknecht 2012 for an overview). I argue that Brentano's reflections can be treated as highlighting one aspect of this question insofar as he not only draws attention to the issue of methodological distinctiveness, but also highlights the difficulties of providing a philosophical justification for it.

Now, beyond such general points about (inter)disciplinarity and integration, can we also gain systematic insights from Brentano's more specific reflections about psychology in its relation to neuroscience? As we saw, Brentano's views about a distinctively psychological method rested on specific assumptions about subject matter and explanatory principles of psychology, i.e., (a) that only conscious mental states could be the subject matter of a "pure", *descriptive psychology*, and (b) that *genetic psychology* would inevitably involve appeal to non-psychological (i.e., physiological) processes. Since neither of these two positions are common in current debates, this raises the question of whether Brentano's views have any relevance today. In the following section, I will address this question, arguing that some of Brentano's views are compatible with more mainstream debates, whereas others fruitfully challenge fundamental assumptions underlying those debates.

6 Continuing Relevance?

The idea that consciousness is the subject matter of psychology was radically questioned by early to mid-twentieth century behaviorism in psychology. This concerned both the very project of providing phenomenological analyses of conscious mental states and the project of explaining behavior by appeal to such states (or any intervening states, for that matter!). With that in mind, it is clear why Brentano's vision of psychology might have seemed rather bizarre to psychologists in the first half of the twentieth century. How has the intellectual landscape changed since then? A first significant shift, surely, was the return of "cognitive" explanations (in germinal form already in certain forms of neo-behaviorism, but certainly with the rise of information-theoretic vocabulary in psychology in the 1950s and 1960s). A second shift was the rise of consciousness–studies as a multidisciplinary field, and with it the return of introspection as a subject of serious philosophical and methodological reflections.

6.1 Questioning the Emphasis on Psychological Explanation

Notice that while the rise of cognitivism was marked by a comeback of appeal to intervening variables as explanatory of behavior, such intervening variables were

typically defined functionally, and were not conceived of as requiring conscious awareness.[6] As such, Brentano's account would have seemed alien to cognitive psychology on two counts: because he viewed mental phenomena as essentially conscious, and, consequently, because he did not believe that an explanation of behavior that appealed to unconscious states was appealing to mental states at all. By contrast, cognitive psychology as it emerged in the 1950s and 1960s (a) viewed mental states or processes as not necessarily conscious (or even consciously accessible), and (b) viewed such states or processes as explanatory in their own right. As such, Brentano's psychology is orthogonal to received views in psychology and cognitive science.

I argue that one important purpose that can be served by seriously considering Brentano's position is that of unsettling received notions in philosophy of psychology. One such received notion is that psychology aims at, and perhaps even succeeds in providing, *explanations*. It is not my aim in this article to argue that this view is false. Rather, my aim is to draw attention to the fact that its truth is not self-evident! Nor is it self-evident what hangs on it. This is brought out when we consider Brentano's argument in its historical and intellectual contexts. I have argued that Brentano's case for methodological distinctiveness has to be seen in the context of boundary work, in the course of which he wanted to provide a rationale both for the internal unity of psychology and for its separateness from neurophysiology. The notion of *boundary work* draws attention to the fact that the lines between scientific disciplines as we know them are the results of historical processes. One thing that the Brentano case brings out is that there are different conceivable ways of drawing the boundaries around psychology, and they come with different visions about the aim and subject matter of psychology. Moreover, each of them is closely tied up with metaphysical accounts of the mental. By describing Brentano as engaged in boundary work, I do not mean to suggest that his philosophical ideas can treated as *mere* responses to boundary quarrels at the time. I do, however, want to use this case to press the question why contemporary philosophers of psychology should be invested in specific ways of drawing the boundaries around psychology.

In the wake of functionalist arguments against mind/brain identity (and the simultaneous rise of cognitive psychology as a field of research), the thesis of explanatory autonomy was very compelling to philosophers of mind and psychology. It is still being pursued by some philosophers of psychology today, typically in relation to debates about multiple realizability (e.g., Aizawa and Gillett 2009), but others have abandoned it in favor of arguments for explanatory integration (e.g., Piccinini and Craver 2011). Both of these approaches seem to agree, however, that the central business of psychology is that of providing *explanations*. Brentano's work draws our philosophical attention to the investigative process as such, specifically raising questions about the methods that do (or should) guide this process. His specific answer, furthermore, brings to the fore the question of whether an argument for the autonomy of different disciplines can be made on the basis of the specific

[6] I have in mind explanatory concepts like encoding, storage, retrieval, etc . . .

methods they use, and what kinds of metaphysical commitments need to be in place to justify the use of these methods. While Brentano himself viewed the distinctiveness of psychological methods as coming in a package with a rejection of purely psychological explanations, it strikes me that the question of methodological distinctiveness and integration can be debated regardless of where one stands with respect to that of explanatory autonomy *vs.* integration.

6.2 First-Person Reports in Cognitive Neuroscience

Let us turn to a currently active field of research where Brentano's methodological views about descriptive psychology are highly relevant, namely the field of cognitive neuroscience, specifically research that attempts to uncover the neural substrates of conscious mental states. I have in mind research that uses brain imaging techniques in conjunction with specific cognitive tasks and first-person reports. This field of research has also rehabilitated as scientifically respectable the question of how introspective (or first-person) reports can figure in this research. It is here that the relevance of Brentano's views is immediately obvious since he explicitly addresses questions about both scope and problems of first-person reports. Both of these topics have received some attention in the recent literature.

Regarding the question of scope, there are two questions: first, whether all phenomenally conscious mental states are accessible to first-person experiences. Second, whether first person reports provide accurate data about phenomenally consciousness mental states. Cutting across these two issues is another question namely whether there are other, more objective, methods of studying subjective phenomenology, either because not all phenomenological states are accessible, or because not all accessible phenomenal mental states can be accurately reported. All of these questions are subject to current debate within current cognitive neuroscience (e.g., Schmicking and Gallagher 2010).

As we saw, Brentano's opinions about these issues were clear and principled: phenomenally conscious mental states (that is, all mental states) were by definition accessible to the first person. However, immediate first-person experiences ("inner perceptions") did not automatically constitute scientifically admissible first-person data. In this vein, his theory of mind made it possible to distinguish conceptually between concurrent and retrospective self-observation. Consequently, he also distinguished between two sources of errors that could occur when employing first-person methods: (1) errors that resulted from the very act of inner observation (it was because of the inevitability of such errors that he ruled out concurrent inner observation as a legitimate method), and (2) errors that could result from faulty recollections of past experiences. He deemed the latter types of error as less fatal and argued for a psychological methodology of retrospective introspection. Finally, with respect to questions about "objective" (behavioural, neurological) indicators of mental states, we have seen that he thought that they could at supplement, but never replace first-person methods.

Let me outline two reasons for my claim that Brentano's views are still relevant here: one concerns the accuracy and one concerns the scope of first-person data. We will begin with the first problem: the literature about how to validate first-person methods has been vexed with the following difficulty: on the one hand, it would seem obvious that it makes sense to correlate different modes of first person access; e.g., immediate and retrospective (Jack and Roepstorff 2003), or direct *vs.* indirect (Overgaard and Sørensen 2004). On the other hand, there is plenty of evidence that points to functional dissociations between the empirical data generated by these different methods (e.g., Marcel 1993). In the light of Brentano's theory of mind, such dissociations are not too surprising. For example, his theory of mind predicts dissociations between data generated by immediate self-observations and data generated by retrospective self-observations. This does not mean that the findings confirm his theory of mind, but it certainly suggests that it might be of value to bring Brentano's specific mix of methodological and theoretical considerations to bear on this topic.

Let us turn to the second issue: the *scope* of first-person methods. Here the question is whether objective (e.g., behavioural) methods provide a means of detecting phenomenally conscious mental states that are not accessible to phenomenal consciousness (Block 2007). That these debates are not mere philosophical fantasies becomes quite clear when we look at some of the scientific literature, dealing with the implications of phenomena like *blindsight* for methodological choices in the study of the mind/brain. Essentially, there are two interpretations of this phenomenon: one reading is that it is possible to make sensory discriminations in the absence of phenomenally conscious sensory experiences. On the other reading, it is possible to have conscious sensory experiences without first-person access to those experiences (See, for example, Busch et al. 2009 vs. Overgaard et al. 2009 for a dispute along those lines, and Lamme 2010 vs. Overgaard 2010). The implication of the latter reading would be that phenomenal sensory experience can be outside the scope of first-person methods. Notice that this position is not dissimilar to one rejected by Brentano in 1874, i.e., that there can be "unconsciously conscious" mental states. And hence, I argue, it is well worth reviewing his specific reasons since they point to the ways in which methodological disagreements can be tied to deeper metaphysical disagreements (cf. Feest 2012a).

7 Conclusion

In this paper I have argued that we read Brentano's philosophy of psychology as (a) arguing for the methodological distinctiveness of psychology, while (b) at the same time suggesting that such distinctiveness was in fact a prerequisite for the integration of psychological and neurophysiological findings in the ongoing process of research. I backed up my thesis by a detailed textual analysis, and I discussed the relevance of the analysis to current issues and debates within the philosophy of psychology.

As I hope to have shown, Brentano's philosophy of psychology is both deeply metaphysical and practice-oriented. As such, I have argued, it has a number of important insights to offer, not only to contemporary philosophical discussions, but also to contemporary research. By construing Brentano's position in the context of boundary debates, I highlighted that it is not self-evident precisely what psychology is and what makes it special vis-a-vis neighbouring disciplines. While it was not my aim to argue for an adoption of Brentano's specific way of drawing the boundaries, I suggested that his emphasis on the question of methodological distinctiveness introduces an important theme into the recent literature about autonomy and integration. Second, I argued that Brentano draws our attention to the fact that psychological *explanation* has received an extraordinary amount of attention, and I have suggested that a shift of focus towards psychological *investigation* might be fruitful. Lastly, I pointed out that Brentano's methodological and metaphysical position is highly relevant to recent and current research in consciousness studies and cognitive neuroscience.[7]

In conclusion, let me emphasize that whereas Brentano specifically focused on first-person reports as an ineliminable component of distinctively psychological methods, I am not committed to the ideas that these are the only distinctive and unique methods psychology has to offer. Thus, I would suggest that his point – to think about distinctive methodological contributions that psychology makes to an interdisciplinary cognitive science – is well taken, regardless of where one stands with respect to Brentano's philosophy of psychology.

References

Aizawa, K., and C. Gillett. 2009. The (multiple) realization of psychological and other properties in the sciences. *Mind and Language* 24: 181–208.

Andersen, H., and S. Wagenknecht. 2012. Epistemic dependence in interdisciplinary groups. *Synthese*. doi:10.1007/s11229-012-0172-1.

Bechtel, W., and R. Richardson. 1993. *Discovering complexity: Decomposition and localization as scientific research strategies*. Princeton: Princeton University Press.

Block, N. 2007. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences* 30: 481–498.

Bordogna, F. 2008. *William James at the boundaries: Philosophy, science, and the geography of knowledge*. Chicago: University of Chicago Press.

Brentano, F. 1874/1973. *Psychologie vom empirischen Standpunkt*. Trans. *Psychology from an Empirical Standpoint*. London: Routledge.

Busch, N., I. Fründ, and C. Herrmann. 2009. Electrophysiological evidence for different types of change detection and change blindness. *Journal of Cognitive Neuroscience* 22: 1852–1869.

Craver, C.F. 2007. *Explaining the brain*. Oxford: Oxford University Press.

[7]I would like to thank Thomas Uebel for inviting me to present this paper at the workshop "New Directions in the Philosophy of Science" (Bertinoro, Italy, October 2012) and for helpful comments on a previous draft.

Feest, U. 2003. Functional analysis and the autonomy of psychology. *Philosophy of Science* 70: 937–948.
Feest, U. 2012a. Introspection as a method and introspection as a feature of consciousness. *Inquiry* 55: 1–16.
Feest, U. 2012b. Husserl's crisis as a crisis of psychology. *Studies in the History and Philosophy of Biological and Biomedical Science* 43: 493–503.
Galison, P., and D. Stump. 1996. *The disunity of science. Boundaries, contexts, and power*. Stanford: Stanford University Press.
Gieryn, T. 1983. Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review* 48: 781–795.
Hatfield, G. 2005. Introspective evidence in psychology. In *Scientific evidence. Philosophical theories and applications*, ed. P. Achinstein, 259–286. Baltimore/London: Johns Hopkins University Press.
Horwicz, A. 1872a. *Psychologische Analysen auf physiologischer Grundlage. Ein Versuch Zur Neubegründung der Seelenlehre*. Halle: C.E.M. Pfeffer.
Horwicz, A. 1872b. Methodologie der Seelenlehre. *Zeitschrift für Philosophie und philosophische Kritik* 60: 163–206.
Jack, A., and A. Roepstorff. 2003. Why trust the subject? *Journal of Consciousness Studies* 10(9–10): v–xx.
Lamme, V. 2010. How neuroscience will change our view of consciousness. *Cognitive Neuroscience* 1: 204–220.
Marcel, A. 1993. Slippage in the unity of consciousness. In *Experimental and theoretical studies of consciousness*, ed. G.R. Bock and J. Marsh, 168–180. Chichester: Wiley.
Maudsley, H. 1867. *Physiology and pathology of the mind*. New York: Appleton.
Mitchell, S. 2003. *Biological complexity and integrative pluralism*. New York: Cambridge University Press.
Overgaard, M. 2010. How consciousness will change our view of neuroscience. *Cognitive Neuroscience* 1: 224–225.
Overgaard, M., and T.A. Sørensen. 2004. Introspection distinct from first-order experiences. *Journal of Consciousness Studies* 11: 77–95.
Overgaard, M., M. Jensen, and K. Sandberg. 2009. Methodological pitfalls in the 'objective' approach to consciousness: Comments on Busch et al. (2009). *Journal of Cognitive Neuroscience* 22: 1901–1902.
Piccinini, G., and C. Craver. 2011. Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183: 283–311.
Schmicking, D., and S. Gallagher (eds.). 2010. *Handbook of phenomenology and cognitive science*. Dordrecht: Springer.
Wundt, W. 1874/1973–1974. *Grundzüge der physiologischen Psychologie*, 2 vols. Reprinted in Leipzig: Engelmann.
Wundt, W. 1888. Selbstbeobachtung und innere Wahrnehmung. *Philosophische Studien* 4: 292–309.

On the Logical Positivists' Philosophy of Psychology: Laying a Legend to Rest

Sean Crawford

1 The Legend

Near the beginning of his recent Carnapian book *Constructing the World*, David Chalmers rightly remarks that "A surprising and often-overlooked feature of the *Aufbau* is that Carnap there requires only that definitions be extensionally adequate" (Chalmers 2012, p. 4). Chalmers then points out that in the 1961 preface to the second edition of the *Aufbau* (Carnap 1928/1961/2003), Carnap "says that the definitions should be held to a stronger, intensional, criterion of adequacy" (Chalmers 2012, p. 5). Chalmers goes on to give "analyticity, apriority and necessity" as examples of stronger criteria (obviously intending the necessity in question to be logical or metaphysical necessity). It is clear from the preface, however, that Carnap did not demand such strongly intensional criteria, but only that the coextensiveness of the *definiendum* and *definiens* not be merely "accidental, but necessary, i.e., it must rest either on the basis of logical rules or on the basis of natural laws" (Carnap 1928/1961/2003, p. ix).[1] Since natural laws are obviously not analytic or knowable a priori, it is clear that Carnap did not think that all philosophical definitions are required to meet the very strong modal criteria of intensionality that Chalmers requires for his constructional project. Hannes Leitgeb (2011, pp. 267, 269–270)

Dedicated to the memory of Rachel Jardine.

[1]Carnap also refers the reader to his reply to Goodman in the Schilpp volume (Carnap 1963) where he says that the identity of extension must fulfill an additional requirement, which "consists in the condition that the correspondence hold, not merely accidentally, but on the basis of general regularities, e.g., physical laws or empirical generalizations" (p. 946).

S. Crawford (✉)
Philosophy, School of Social Sciences, University of Manchester,
Oxford Road, Manchester M13 9PL, UK
e-mail: sean.crawford@manchester.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_49,

also draws attention to the extra demand stated by Carnap in the preface, but, unlike Chalmers, notes explicitly that the intensionality in question may be merely one of nomological necessity. However, it is important to note – and this something that neither Chalmers nor Leitgeb mention, as they are concerned solely with exploring the possibility of actually carrying off a suitably revised *Aufbau*-like construction project – that the stronger requirement of some degree of intensional adequacy for definitions is acknowledged by Carnap very soon after the *Aufbau*.

Just such an intensional adequacy condition on definitions short of analyticity (and hence a priority) finds expression, for example, four years later in *Unity of Science* (Carnap 1932a/1934). One of the aims of that work is to show that protocol sentences about subjective experience can be physicalized, that is, translated into sentences employing only physical vocabulary (which is the only vocabulary known to be inter-subjective, inter-sensory, and universal, which are the requirements for any suitable confirmation base for empirical science). It is evident that Carnap there envisages such "translations" to be based partly on natural laws and hence neither logically necessary nor analytically true nor knowable a priori. The same is true of the broader goal of physicalizing the science of psychology (one element of unified science). The proposed physical definitions of psychological terms and physical translations of psychological sentences are required to meet (to borrow Chalmers' apt phrase) stronger-than-extensional conditions of adequacy – but, contrary to Chalmers's own constructional requirements, only up to *nomological* intensionality. That is to say, the proposal is that it is only nomologically necessary that the physical *definiens* and the psychological *definiendum* have the same extension, not logically necessary (let alone that the two be synonymous).

Just as the original extensionality of the *Aufbau* construction is often overlooked,[2] it is a virtually universally overlooked feature of Carnap's philosophy of psychology of the 1930s – the period in which he is alleged to have been a logical behaviourist – that his proposed physical-behavioural definitions of mental terms are only required to be intensionally adequate up to nomologicality – which means, contrary to popular belief, that Carnap was not a logical behaviourist. In so far as Hempel and fellow logical positivists agreed with Carnap on this, which they certainly seemed to, an even more general claim is warranted, namely, that the logical positivists were not logical behaviourists.

Yet Carnap and Hempel and their fellow logical positivists are considered to be examples of logical behaviourists *par excellence*. For example, in his celebrated 1963 critique of logical behaviourism, "Brains and Behaviour", Hilary Putnam wrote that

> the Vienna positivists in their 'physicalist' phase (about 1930) … [produced] the doctrine we are calling *logical behaviourism* – the doctrine that, just as numbers are (allegedly) logical constructions out of *sets*, so *mental events* are logical constructions out of actual and possible *behaviour events*. (Putnam 1963, p. 326)

[2]Even by Quine, to some extent, for as Nelson Goodman (1963, p. 555n5) has pointed out, "The avowed extensionalism of so outstanding a monument of phenomenalism and constructionism as the *Aufbau* would seem to confute Quine's recent charge [in 'Two Dogmas'] that the notion of analyticity is a 'holdover of phenomenalistic reductionism'."

Logical behaviourism so understood was an "extreme thesis" implying that "all talk about mental events is translatable into talk about actual or overt potential behaviour" (*ibid*.). Putnam then claimed that "In the last thirty years, the original extreme thesis of logical behaviourism has gradually been weakened to something" which, he said, "a great many" philosophers at the time would have accepted. The central tenet of the weakened thesis is that "[t]here exist entailments between mind-statements and behaviour-statements; entailments that are not perhaps analytic in the way that "All bachelors are married" is analytic, but that nevertheless follow (in some sense) from the meanings of mind words" (*id*., p. 327). Putnam proposed to call these entailments "*analytic entailments*."[3] He eventually went on, of course, to offer the undergraduate philosophy student's favourite refutation of logical behaviourism: the fictional community of "super-spartans" (or "super-stoics"). These legendary people, as everyone knows, are capable of suppressing all pain behaviour even while suffering "the agonies of the damned" – thus supposedly demonstrating that the connection between statements about pain and statements about pain behaviour cannot be a matter of analytic meaning equivalence or even one-way analytic entailment of any kind. The connection between the two is at best entirely synthetic – or, as we would put it nowadays, contingent.

As anyone familiar with philosophy of mind textbooks and anthology introductions can attest to, Putnam's understanding of the logical positivists' logical behaviourism is the orthodox one in the analytic tradition and continues to be so. Forty years later, for example, we find John Searle characterizing logical behaviourism as "a logical thesis about the definition of mental concepts. ... The model for the behaviourists was one of definitional identities. Pains are dispositions to behaviour in a way that triangles are three-sided plane figures. In each case it is a matter of definition" (2004, pp. 54–55), and citing C. G. Hempel as an example of a philosopher endorsing this kind of behaviourism.[4] Since the "definitional identities" Searle refers to, as well as Putnam's "analytic entailments", are supposed to be knowable a priori, it follows that the analyses or translations or entailments proposed by logical behaviourists, as holding between psychological statements and behavioural statements, are supposed to be knowable a priori. We can, then, sum up the received view in analytic philosophy of the logical positivists' logical behaviourism as the thesis that the relation between psychological statements and the behavioural statements intended to give their meaning is an a priori knowable analytic entailment or equivalence. This is of course why logical behaviourism is also known as "analytical behaviourism."

As I have already suggested, when one turns to the two canonical logical positivist texts of logical behaviourism, one finds a very different story. Neither Rudolf Carnap's "Psychology in Physical Language" (1932) nor C. G. Hempel's "The Logical Analysis of Psychology" (1935) espouses the received view of logical behaviourism just adumbrated; nor do any of their subsequent writings

[3]Cf. (Putnam 1969).

[4]Soon after Putnam, similar claims were be made by Fodor (1968, pp. 51, 155n6), Cornman (1971, pp. 132ff, esp. p. 140), and Kim (1971, p. 328).

on philosophy of psychology. First, neither Carnap nor Hempel restrict their reductive analyses of psychological statements to statements describing *behaviour*; on the contrary, they both included in their analyses reference to internal neurophysiological states. Secondly, in the eyes of both Carnap and Hempel, most of the logico-semantic relations between psychological statements and both the behavioural statements and the physical statements intended to give their meanings are synthetic and knowable only a posteriori.

Let us call the historical claim that the logical positivists maintained that the connections between "mind talk" and "behaviour talk", or between "mind talk" and "physical-thing talk", were analytic, the *analytic entailment interpretation* of logical positivism's logical behaviourism. The fact is that the analytic entailment interpretation is as much a fiction as Putnam's super-spartans. But if so, how did this extraordinary and ubiquitous legend get perpetuated throughout the analytic tradition? My chief aim here is to try to make some progress toward answering this question. I shall first establish (in Sect. 2) that the analytic entailment interpretation is indeed nothing more than a legend.[5] Second (in Sect. 3), I will make some conjectures about why it became so widely accepted.

2 Laying the Legend to Rest

In his Physicalist writings of the 1930s Carnap refers to "rules of transformation" (sometimes "rules of inference" and "rules of translation") of the physical language in which definitions and translations are to be carried out in accordance with the programme of Unified Science. Although he is not always entirely explicit about it, it is evident to a careful reader that even in the earliest of these writings from his Physicalist phase that not all of these inference rules are laws of logic and that some of them are intended to be laws of nature. For example, in "Unity of Science" (1932a/1934) Carnap writes of the "the rules of transformation inside the physical language (*including the system of natural laws*)" (p. 88, my emphasis; cf. p. 92). This becomes much clearer by the time of *The Logical Syntax of Language* (1934) and "Testability and Meaning" (1936–1937) in which Carnap explicitly distinguishes between the L-rules and the P-rules of a scientific language on the basis of which transformations may be validly carried out: the former are logical laws and the latter empirical physical laws. (Carnap also defines various correlative notions, such as L-validity and P-validity, L-equipollence and P-equipollence, and L- and P-synonymy.) Both kinds of "translation rules" are to be employed in physicalization.

[5]The only writers I know of who explicitly do not hold the analytic entailment interpretation are Alston and Nakhnikian (1963, p. 391), Hempel (1958, 1969), Cirera (1993) and Kim (2003). I am grateful to my colleague Thomas Uebel for drawing my attention to Cirera's important article as well as to Hempel (1969) and much other relevant literature, as well as for many edifying discussions of Carnap and logical positivism, which prompted several changes to this chapter.

In *Philosophy and Logical Syntax* (1935/1963), Carnap claims that "every sentence of any branch of scientific language is equipollent to some sentence of the physical language, and can therefore be translated into the physical language without changing its content" (p. 455; cf. *The Logical Syntax of Language* §82). Carnap is very clear in this work (as well as in *The Logical Syntax of Language* §§51, 82) that there can be two concepts of equipollence, that is, equivalence, in the physical language: logical equipollence (L-equipollence) and physical equipollence (P-equipollence). Two sentences are L-equipollent when they are mutually derivable solely on the basis of logical laws; two sentences are P-equipollent when they are mutually derivable only on the basis of physical laws as well.

Given this distinction between L- and P-concepts, in order for the legendary analytic entailment interpretation to be true, Carnap would have to maintain that all physical translations of psychological sentences are L-equipollent to them. But he did not maintain this. He explicitly allowed a psychological sentence, Q_1, and a physical translation of it, Q_2, to be P-equivalent, as Q_1 could be transformed into Q_2 on the basis of "a scientific law, that is, a universal sentence belonging to the valid sentences of the scientific language-system" (1935/1963, p. 456). He took some pains to point out that, in his view, this universal sentence "need not be analytic; the only assumption is that it is valid. It may be synthetic, in which case it is P-valid" (*ibid.*).[6] Similarly, with respect to the translation of the psychological statement "Mr. A is now excited" (P_1), Carnap (1932b/1959) says that it

> cannot, indeed, today be translated into a physical sentence P_3 of the form "such and such a physico-chemical process is now taking place in A's body" (expressed by a specification of physical state-coordinates and by chemical formulae). Our current knowledge of physiology is not adequate for this purpose. (1932b/1959, p. 175)

[6]In a letter he wrote to Herbert Feigl in 1933 (translated by and quoted in Feigl 1963, p. 255), Carnap explicitly states that the two sentences are not analytic. He offers two translations of "N. has a visual image of a house" (A), "The organism of N. is in the state of house-imagining" (B_1) and "In the organism of N. there is an electrochemical condition of such a kind (described in terms of electrochemistry)" (B_2), and then remarks that:

> Both B_1 and B_2 are translations of A. According to my recently adopted terminology, I assert: A is equivalent ("*gehaltgleich*") to both statements ... ; viz., L-equivalent (*logically* equivalent) with B_1; but P-equivalent (*physically* equivalent) with B_2, i.e., mutually translatable (derivable) using besides the logical laws also natural laws as rules of inference, incorporated as transformation rules in the scientific language. You are therefore right in saying that B_2 is only synthetically equivalent with A.

As Ramon Cirera (1993) importantly points out, while B_1, unlike B_2, is claimed by Carnap to be L-equivalent to A, it is not behavioural – in fact, it is not even physical. Neither Feigl nor Cirera say what the point of B_1 is. One possibility is that it is an adverbial analysis of (A) intended to avoid commitment to the intentional object apparently designated by the phrase "visual image of a house", and hence to avoid intentional language, thus making the ultimate goal of a physical translation into B_2 easier. Such adverbial techniques were sometimes employed by Russell in order to avoid commitment to intentional objects (and by some of the American New Realists in a quasi-behaviourist spirit) and Russell of course influenced Carnap. Chisholm (1955–1956) famously criticized such adverbial strategies for avoiding intentional language but I know of no response by Carnap to Chisholm on this point.

It goes without saying that Carnap did not view knowledge of physiology as a priori. It is lack of a posteriori knowledge of physiology that prevents P_1 from being translated into P_3. It follows, of course, that the relation between P_1 and P_3 is synthetic, not analytic.

It might be objected that it is simply not true that, in order for the analytic entailment thesis to hold, all physical translations of psychological sentences must be L-equivalent to them. It is only required that the physical-*behavioural* ones be L- equivalent. What Carnap has in mind here is the case in which Q_2 is a non-behavioural, neurophysiological sentence. This of course can only be P-equivalent. But that is entirely consistent with claiming that a behavioural sentence, Q_3, must be L-equivalent with the psychological sentence of which it is a translation. Moreover, this fits nicely with the logical positivists' view (emphasized by Feigl (1958)) that the progress of the science of psychology will unfold in two stages: an initial "peripheralist" black-box behaviouristic stage and a later "centralist" neurophysiological stage.

Carnap does indeed countenance physicalizations that are explicitly only about overt behaviour. In "Psychology in Physical Language", he says that P_1 ("Mr. A is now excited") may be inferred from p_1, which is a sentence "about the behaviour of A, e.g. about his facial expressions, his gestures, etc. or about physical effects of A's behaviour, e.g. about characteristics of his handwriting" (*id*., p. 171). The question is whether Carnap intends P_1 and p_1 to be L-equivalent, as the analytic entailment interpretation requires, or whether they are supposed to be merely P-equivalent – just as P_1 and P_3 are. Although he does not employ L- and P-concepts explicitly in "Psychology", it is nonetheless clear that the logical relation between P_1 and p_1 is synthetic and therefore that the two are intended to be at best P-equivalent.[7] For he tells us that P_1 is to be derived from p_1 on the basis of the "major premise O", which states that "When I perceive a person to have this facial expression and hand-writing he (usually) turns out to be excited. (A sentence about the expressional or graphological signs of excitement.)" (*ibid*.). Earlier in the paper, Carnap draws a distinction between singular and general scientific – including psychological – sentences (*id*., p. 168), taking pains to emphasize (against what he calls "phenomenology" and its objectionable a priori or at least non-inductive methods) that general sentences are discovered inductively and are therefore to be considered empirical hypotheses (*ibid*.). He then explicitly says that the major premise O, on the basis of which P_1 is derived from p_1, is such a general sentence (*id*., p. 171). O's synthetic status (as an empirical inductive hypothesis) is therefore beyond question. It follows that the connection between P_1 and p_1 is synthetic too; generalizing, the connection between psychological sentences and their behavioural protocols is, contrary to the analytic entailment interpretation, synthetic.

[7] And perhaps not even that, owing to the fact that the connection between the two may not even be nomological, as we shall see presently.

This is further highlighted by what Carnap goes on to say:

> The cited relationship between P_1 and p_1 may also be seen in the fact that under certain circumstances, the inference from p_1 to P_1 may go astray. It may happen that, although p_1 occurs in a protocol, I am obliged, on the grounds of further protocols, to retract the established system sentence P_1. I would then say something like, 'I made a mistake. The test has shown that A was not excited, even though his face had such and such an expression'. (*ibid.*)

Similarly, 5 years later in "Logical Foundations of the Unity of Science", he writes:

> Let us take as an example the term 'angry'. If for anger we knew a sufficient and necessary criterion to be found by a physiological analysis of the nervous system or other organs, then we could define 'angry' in terms of the biological language. The same holds if we knew such a criterion to be determined by the observation of the overt, external behaviour. But a physiological criterion is not yet known. And the peripheral symptoms known are presumably not necessary criteria because it might be that a person of strong self-control is able to suppress these symptoms. If this is the case, the term 'angry' is, at least at the present time, not definable in terms of the biological language. But, nevertheless, it is reducible to such terms. ... The logical nature of the psychological terms becomes clear by an analogy with those physical terms which are introduced by reduction statements of the conditional form. Terms of both kinds designate a state characterized by the disposition to certain reactions. In both cases the state is not the same as those reactions. Anger is not the same as the movements by which an angry organism reacts to the conditions in his environment, just as the state of being electrically charged is not the same as the process of attracting other bodies. In both cases, that state sometimes occurs without these events which are observable from outside; they are consequences of the state according to certain laws and may therefore under suitable circumstances be taken as symptoms for it; but they are not identical with it. (Carnap 1938, pp. 56–57, 59)[8]

We shall return to Carnap's important distinction between definition and reduction below (in Sect. 3) because it is highly relevant to explaining the origins of the legend. For the moment, however, let us note that Carnap here anticipates and preempts Putnam's super-spartan objection by more than two decades, explicitly stating that he does not view the behavioural dispositions associated with anger as identical with it; on the contrary, the behavioural "symptoms" of anger are (normally) caused

[8]Hempel's discussion in "The Logical Analysis of Psychology" (1935, esp. §V) is considerably less clear about this, and this lack of clarity may well have contributed to the legend, especially given that Hempel's article is more widely reprinted than Carnap's "Psychology" (it appears, e.g., in the highly influential collection *Readings in Philosophical Analysis* (Feigl and Sellars 1949)). Hempel there confusingly claims that it is *logically contradictory* to say that all the symptoms obtain but the psychological state does not. This seems to be because, first, unlike Carnap, he is using the term "symptoms" (sometimes putting it between inverted commas) to cover not only the external behavior but also the internal physiological processes associated with the psychological state, and, second, he is heading off a dualist objection. Still, given that Hempel maintains that all these "symptoms" are discovered empirically (cf. note 21 below), he cannot really mean that their presence with the absence of the psychological state is logically contradictory. Rather, the sentence describing such a situation would be (at best) what Carnap (1934, §52) calls *P-contravalid* (i.e., nomologically impossible). Hempel's confusion here may be of a piece with the one Feigl and others make about the nature of definition, as discussed below in §3.

by anger. This is already clear in 'Psychology' for, as we saw above, he considers the general sentence O to state behavioural "signs" of excitement – in other words, "symptoms", that is, causal effects of excitement, noting that they may be present even when excitement is not.[9] In this he also anticipates the later causal critiques of behaviourism levelled by Jerry Fodor and David Armstrong (that mental events are not identical with behaviour but are the causes of behaviour). The central point, again, is the fact that even here, with an explicitly overt-behavioural proposal, Carnap does not declare that such an overt-behaviouristic reduction sentence for anger will be analytic. The "laws" referred to, connecting inner states with outer behavioural reactions or symptoms, are empirical physical laws and so are intended to be P-valid (at best) and so synthetic.

Jaegwon Kim (2003, p. 275), however, sees Carnap as anticipating a causal-functional analysis of mental concepts and so as lending support to the idea that while Carnap may have viewed the nomological correlation of psychological sentences with neurophysiological sentences as P-equivalent, he viewed the correlation of psychological sentences with *behavioural* sentences as L-equivalent and hence analytic. Now, it is true, I think, that Carnap does here anticipate – remarkably – a kind of functionalism. However, contrary to what Kim says, it is, I think, incorrect to associate Carnap with the *analytic* functionalism of David Lewis and others, which is really a development of Rylean ordinary-language logical/analytical behaviourism. According to analytic functionalism, the causal-functional definitions of mental terms are specified a priori by conceptual analysis of commonsense psychology, and the only role for empirical science is to discover a posteriori which inner states are the actual physical realizers of the definitions. Carnap's proto-functionalism is, I believe, more akin to a thorough-going empirical psycho-functionalism (Block 1978), in which empirical science is involved at the first stage; that is, the functional definitions of mental terms are themselves in many cases specified empirically by scientific investigation.

In a word, Carnap's proto-functionalism is what one might call *synthetic functionalism*. This fits the text and the spirit of "Psychology" and "Foundations" better, to make better sense of the strong analogy Carnap draws between concept formation in the natural sciences and in the sciences of psychology, and to gel better with Carnap's (1932b) procedure of physicalization.[10] This was certainly Hempel's view of Carnap, who maintained that for the latter "those behavioural symptoms which are generally associated with a given psychological feature will often be determined by empirical investigations leading to empirical laws rather than by

[9]The inner event of excitement that is the cause of the outer behavioural symptom is eventually to be identified with the inner neurophysiological state that is the cause of the behavioural symptom, *à la* later causal-role functionalism. See immediately below for more on this.

[10]I discuss Carnap's empirical procedure of physicalization in slightly more detail in Crawford (2013).

an aprioristic reflection upon the meaning of the psychological terms in question" (Hempel 1969, pp. 179–180).[11]

In short, Carnap went the whole hog: most – perhaps all – of the semantic relations between *both* psychological sentences and *neurophysiological* sentences, and psychological sentences and physical-*behavioural* sentences, were synthetic.

Indeed, there is strong reason to doubt that in the period with which we are concerned – circa 1930–1940 – Carnap could even have held onto *any* analytic entailments between psychological and behavioural statements, even if he had wanted to. To see this, we need only remember the distinction between the narrow and the broad notion of analyticity, famously emphasized by Quine (1951) in "Two Dogmas". In the narrow sense, an analytic truth is a logical truth, that is, a truth based solely on the meanings of the logical constants, such as the tautology "If p then p"; such a logical truth remains true under all "re-interpretations" of the meanings of its non-logical terms (such logical truths are now more commonly called *logically valid* sentences or formulae). An analytic truth in the broad sense is a truth based solely on the meanings of the logical terms and the non-logical or "descriptive" terms; an "all-bachelors-are-unmarried" kind of truth, as it were. The analyticity of these latter "truths of essential predication", as Quine (1963) calls them (apparently after Mill), lies in the fact that they can be transformed into logical truths by substitution of definitions or synonymies. It is, of course, the broad notion of analyticity that is at stake here; no one ever claimed that logical behaviourists thought the entailments between psychological and physical-behavioural sentences were tautological or trivially logically true. But this crucial distinction between the broad and the narrow senses of analyticity seems to have occurred to Carnap only after 1940 and some late remarks by him (Carnap 1964/1994, p. 259) seem to suggest that it was Quine who woke him from his dogmatic Wittgensteinian slumbers (according to which analyticity just is logical truth) that he began to

[11]Moreover, as Carnap himself later pointed out (1952, p. 71) – and as is noted by Hempel (1951, p. 72; 1952, p. 28) and Arthur Pap (1958, Chap. 11) – if there is more than one partial or conditional definition, e.g., a pair of (either unilateral or bilateral) reduction sentences for a given term, as obviously Carnap expected their to be for theoretical terms of behavioural psychology, then one can derive a synthetic statement from them, from which it follows that at least one of the definitions must be synthetic. See Carnap (1936–1937) for the notion of a reduction sentence. I discuss the difference between reduction and definition in §3 below. Carnap (1952) ingeniously goes on to suggest a procedure to overcome the fact that pairs of reduction sentences introducing a theoretical term will have synthetic consequences, by taking a weaker (material) conditional sentence, whose antecedent is a statement of the empirical content of the reduction pair (the "representative sentence", as he (1936–1937) called it) and whose consequent is the reduction pair, as the "meaning postulate" introducing the theoretical term, because none of its logical consequences containing only the original defining (basic) terms are synthetic. However, it is important to note for present purposes that while none of these logical consequences are synthetic, they are analytic only in the narrow sense, that is, they are logical truths. Such meaning postulates cannot therefore underwrite behavioural definitions for psychological terms in the spirit of textbook logical/analytical behaviourism, which obviously requires a broader notion of analyticity. See immediately below for further relevant discussion of this point.

distinguish explicitly between logical truth and analyticity in the broader sense and attempt to define the latter as well as the former.[12]

I cannot be wholly definite here. The received view has it that it was only later, spurred by Quine's attack on the analytic/synthetic distinction in "Two Dogmas" and his drawing there of the distinction between logical truth in the narrow sense and analyticity in the broader all-bachelors-are-married sense, that Carnap undertook the task of characterizing analyticity in the broader sense. Accordingly, Carnap first proposed to define the broader sense of analyticity for an artificial observational language in "Meaning Postulates" (Carnap 1952), which is repeated in chapter 27 of *An Introduction to the Philosophy of Science* (Carnap (1964/1994), and for an artificial theoretical language in Chap. 28 of that same work; and that his proposal for natural languages first appears in "Meaning and Synonymy in Natural Languages" (Carnap 1955; see also Carnap's reply to Hempel in the Schilpp volume).[13] Be that as it may, however, it suffices for the present argument that during the time in which Carnap formed his views on the philosophy of psychology under discussion here (roughly, the late 1920s to the late 1930s) he did equate analyticity with logical truth and so could not have endorsed textbook logical behaviourism.

3 Origins of the Legend

So how on earth did the analytic entailment interpretation ever get started in the first place?

One source of the legend, which space limitations prevent me from exploring in any detail, is probably the often-drawn analogy with phenomenalism and logical constructionism. We have already seen Putnam claim that logical behaviourism was a logical constructionist thesis akin to the construction of numbers out of sets – and obviously the latter construction will involve exclusive use of L-rules. Similarly, if one thinks of phenomenalism as the doctrine that material-object statements analytically entail sense-data statements – as for example in Chisholm (1957, Appendix) – and one thinks logical behaviourism is like phenomenalism, then one

[12] Actually, as Quine points out in "Two Dogmas" (§1, p. 41), even in the modal-logic phase of *Meaning and Necessity* (Carnap 1947) in his later "semantic period", when he defined analytic truth semantically as truth in all state descriptions, Carnap's definition was still only of the narrower notion of analyticity as logical truth.

[13] Awodey (2007, p. 244n 30) endorses this take on the matter, which does indeed seem to be Carnap's own view, at least in some of his later writings (e.g., Carnap 1964/1994, p. 259). It is absolutely clear, however, despite what Carnap says in these writings, that he was alive to the importance of the distinction between narrow logical truth and broad analyticity as far back as 1943 (and probably earlier). See, e.g., the letter Carnap wrote to Quine on 21 January, 1943 (printed in Creath 1992, pp. 303ff). The nature and development of Carnap's views on how to define formally the difference between narrow analyticity as logical truth and broad analyticity as essential predication has not to my knowledge been studied in detail let alone resolved in a fully satisfying manner. I hope to discuss it in future work.

will no doubt arrive at the analytic entailment interpretation. But as we have seen, it is simply a mistake to view the logical positivists' logical behaviourism as like phenomenalism, or more generally logical constructionism, in this sense.[14]

More importantly, we can trace the origin of the analytic entailment interpretation largely to the positivists' highly technical and idiosyncratic use of the expressions "translation", "meaning", "synonymy", "definition" and their cognates. None of these terms is used today in anything like the way the logical positivists, especially Carnap and Hempel, were using them in the 1930s and even to some extent in the 1940s.[15] Most of these terms and their cognates have strong modal implications for us now that they did not have back then for the positivists, namely a degree of intensionality up to at least logical necessity and perhaps even hyperintensionality. Carnap and Hempel, however, were working with a background extensional logic. When they claim that "mind talk" can be *translated* into "physical talk", what they mean is that one can construct material bi-conditionals with mind talk on the left-hand side and physical-thing-language talk on the right-hand side. These material bi-conditional "translations" were just that – *material* bi-conditionals, containing the straightforward truth functional connective symbolized by the horseshoe. Most of these material bi-conditionals (and reduction sentences) were clearly understood at the outset to be synthetic statements of lawful correlations discovered empirically through scientific investigation. Carnap (1956) is especially clear about this. Of course, philosophers of science, including Carnap himself (1936–1937), soon began to realize the extreme difficulties encountered in formalizing natural laws, disposition statements, and the counterfactual conditionals associated with them using an extensional logic.[16] In light of this, Hempel remarks on one alternative approach that appeals to empirical causal laws: "The extensional 'if … then …' – which requires neither logical nor nomological necessity of connection – would therefore have to be replaced … by a stricter, nomological counterpart which might be worded perhaps as 'if … then, with causal necessity, …'" (Hempel 1958, p. 188).

The most significant instance of this confusion of extensional and intensional semantic concepts as applied to the philosophy of psychology seems to occur in the work of the ambivalent and erstwhile logical positivist Herbert Feigl; and I conjecture that it is perpetuated and carried into contemporary analytic philosophy of mind's self-image through the Feigl-Putnam-Fodor line of influence. Early on in his famous long essay, "The 'Mental' and the 'Physical'", Feigl explains the transition from logical behaviourism to the mind-brain identity theory:

---

[14]Unless – ironically – one is explicitly thinking of *Carnap's* (as opposed, e.g., to A.J. Ayer's) phenomenalism and logical constructionism. See note 2 above. So while there is a parallel between *Carnap's* phenomenalism/logical constructionism and logical behaviourism, it is precisely the opposite of what that parallel is usually taken to be. Carnap's phenomenalism and his behaviourism were *both* synthetic.

[15]Ducasse (1941, ch. 7) complains that what Carnap (1935/1963) calls translation is not truly translation.

[16]See Carnap (1956), Hempel (1958) and Suppe (1977) for discussion of this.

> A most important *logical* requirement for the analysis of the mind-body problem is the recognition of the *synthetic* or *empirical* character of the statements regarding the correlation of psychological to neuro-physiological states. It has been pointed out time and again that the early reductionistic logical behaviorism failed to produce an adequate and plausible construal of mentalistic concepts by explicit definition on the basis of purely *behavioral* concepts. ... I was tempted to identify, in the sense of *logical* identity, the mental with the neurophysiological ...
>
> But if this theory is understood as holding a *logical translatability* (analytic transformability) of statements in the one language into statements in the other, this will certainly not do. ...
>
> [T]he question which mental states correspond to which cerebral states is in *some* sense ... an empirical question. If this were not so, the intriguing and very unfinished science of psychophysiology could be pursued and completed by purely a priori reasoning. ...
>
> ... Subjective experience ... cannot be *logically* identical with states of the organism; i.e., phenomenal terms could not explicitly be defined on the basis of physical$_1$ or physical$_2$ terms. (1958, pp. 389–390)

Aside from encouraging the erroneous view that early reductionistic logical behaviourism was purely overt-behavioural, excluding reference to inner neurophysiological states, while his own early view (Feigl 1950) included them, Feigl runs together two crucially different things: *analyticity* and *definability*.[17] He assumes that an explicit definition cannot be synthetic but can only be analytic and consequently assumes that abandoning the idea of explicit definition is tantamount to embracing the idea that the connection between what was originally the *definiendum* and *definiens* is synthetic.[18] But these assumptions are mistaken.

According to Carnap and Hempel, if a non-primitive expression, the *definiendum*, is explicitly definable in terms of primitive expressions, then it can be eliminated and replaced by its *definiens*, by the primitive expressions. Such explicit definitions were understood by Carnap and Hempel to be the specification of necessary and sufficient conditions for the *definiendum*; that is, the construction of a material bi-conditions whose right-hand side, the *definiens*, contains only undefined primitive terms.[19] For the logical positivists, of course, the defined expressions – in our case, mental ones – will be so-called "theoretical" terms and the primitive expressions the "observation" terms – in our case, physical-behavioural ones. Now, Carnap (1936–1937) very early on saw that the search for explicit definitions of all empirical scientific terms in the physical-thing language, on the basis of which physical translations could be carried out, was misconceived – especially in the case of dispositional terms – and consequently weakened the project to one of providing what he called "reduction sentences", which were either material conditionals with further material

[17] It must be conceded that Hempel (1935, §V) may not have been altogether free of this conflation either. See note 8 above.

[18] Cf. Feigl (1958), pp. 427, 447, as well as Feigl (1963), p. 251 and Feigl (1971), p. 302. Pap (1952, p. 210) and Laurence D. Smith (1986, p. 53) also seem to hold these mistaken assumptions.

[19] Strictly speaking, only the ultimate definition in a definition chain will have only undefined primitive terms in the *definiens*, but this does not affect the present point. See Carnap (1936–1937) and Hempel (1952).

conditionals as consequents or material conditionals with material bi-conditionals as consequents.[20] These reduction-sentence conditionals linked the empirical term in question to physical conditions only under certain test circumstances. Since these physical reduction sentences were not definitions of the terms they were reducing – they were only incomplete "conditional definitions" – they did not allow the terms to be eliminated and replaced and hence they could not form the basis for translations.[21]

The important point for present purposes about this shift from definition to reduction or partial definition is that, with respect to the physicalization of psychology and other empirical sciences, it is not a shift from the category of analytic truths knowable only a priori to the category of synthetic truths knowable only a posteriori. Rather, it is a shift *within* the single category of synthetic truths knowable only a posteriori from complete definability (which permits elimination of the defined term) to incomplete or conditional definability (which does not permit elimination of the partially defined term). Contrary to what Feigl and others seem to suppose, the failure of explicit definition and translation is not at all tantamount to the failure of a priori analytic definition and translation. As we saw in Sect. 2, there never was any such latter project for psychology or indeed any empirical science.

The mistaken assumption that explicit definitions are all and only strongly intensional up to analyticity seems to have been abetted by misinterpretations of the addenda that Carnap and Hempel added to later reprintings of their respective articles. The addenda to Carnap (1932b) and Hempel (1935) state that the two philosophers no longer hold the strict definability thesis and have replaced it with the more flexible reducibility thesis.[22] In his 1977 "prefatory note" to the reprinting in Block (1980), Hempel tells us that he had reservations about agreeing to the reprinting because he no longer held the "narrow translationist form of physicalism [there] set forth" but "yielded to Dr. Block's plea that it offers a concise account of an early version of logical behaviourism" (p. 14). On Kim's interpretation, this implies that "Hempel was in agreement with Block's assessment that logical

[20] See Carnap (1936–1937), §10 and Hempel (1952).

[21] Moreover, as Hempel makes clear, even if necessary and sufficient observational conditions for a theoretical term could be discovered inductively for a merely partially defined theoretical term introduced by reduction sentences, the bi-conditional representing this finding, "Q iff O", where "Q" is the theoretical term and "O" the observational one, "clearly does not express a synonymy; if it did, no empirical investigations would be needed in the first place to establish it. Rather, it states that, as a matter of empirical fact, 'O' is co-extensive with 'Q', or, that O is an empirically necessary and sufficient condition for Q" (Hempel 1958, p. 192). But see note 8 above.

[22] See Carnap's 1957 addendum to the reprinting of Carnap (1932b) in Ayer (1959), Carnap's 1961 addenda to the reprinting of Carnap (1935) in Alston and Nakhnikian (1963), Carnap's preface to the second edition of the *Aufbau* (Carnap 1928/1961/2003), and Hempel's 1972 "Author's preamble" to the reprinting of Hempel (1935) in Marras (1972, p. 115), which are all but identical to Hempel's 1977 "Author's prefatory note" to the reprinting in Block (1980). As Hempel notes in these addenda, physicalization was liberalized even further with the later introduction of "hypothetical constructs" connected to the observation language via "correspondence rules." See also Carnap (1956) and Hempel (1951, 1952, 1958).

behaviourism was the position advocated in his 1935 paper" (2003, p. 266). Since Kim understands logical behaviourism as the thesis that psychological sentences analytically entail physical-behavioural sentences – in other words, since Kim holds the analytic entailment interpretation – he is claiming that Hempel is implying that he (Hempel) advocated the latter thesis in his original article. Kim goes on to point out how problematic Hempel's note is *so interpreted* because hardly any of Hempel's (1935) proffered physical-behavioural conditions are analytically entailed by his sample psychological sentence "Paul has a toothache".[23] But there is no such implication. There is absolutely nothing in Hempel's note to suggest he understood early logical behaviourism as the thesis that psychological sentences analytically entail physical-behavioural sentences. Kim's interpretation can be arrived at only on the assumption that explicit definitions are analytic. But Hempel was never under any such illusion. On the contrary, he is clear that he understands his early version of logical behaviourism to be the claim that psychological concepts are explicitly definable in physical terms and his point is that he has now moved to the more liberal thesis of reduction. Both the earlier definitions and the later reductions were synthetic.[24] Carnap's addendum makes exactly the same point.[25]

Echoing Donald Davidson's famous remark about language, I think we may conclude that there is no such thing as the logical behaviourism of the logical positivists, not if logical behaviourism is anything like what most philosophers have supposed.

References

Achinstein, P., and S.F. Barker (eds.). 1969. *The legacy of logical positivism*. Baltimore: The John Hopkins Press.

Alston, W., and G. Nakhnikian (eds.). 1963. *Readings in twentieth century philosophy*. New York: The Free Press of Glencoe.

Awodey, S. 2007. Carnap's quest for analyticity: the studies in semantics. In *The Cambridge companion to Carnap*, ed. M. Friedman and R. Creath. Cambridge: Cambridge University Press.

Ayer, A.J. (ed.). 1959. *Logical positivism*. New York: Glencoe.

[23] See Crawford (2013) for more detail.

[24] Although, again, as discussed in note 8 above, Hempel (1935) is admittedly not entirely clear about this.

[25] Moreover, to come full circle, the first of the two main changes Carnap announces in the preface to the second edition of the *Aufbau* (the second being the one discussed by Chalmers and Leitgeb which I mentioned at the outset, namely, the shift from extensionality to either logical or nomological intensionality) is the "realization that the reduction of higher level concepts to lower level ones cannot always take the form of explicit definitions. ... The positivist thesis of the reducibility of thing concepts to autopsychological concepts remains valid, but the assertion that the former can be defined in terms of the latter must now be given up and hence also the assertion that all statements can be translated into statement about sense data. Analogous considerations hold for the physicalist thesis of the reducibility of scientific concepts to thing concepts and the reducibility of heteropsychological concepts to thing concepts" (Carnap 1928/1961/2003).

Block, N. 1978. Troubles with functionalism. In *Minnesota studies in the philosophy of science*, vol. 9, ed. C.W. Savage. Minneapolis: University of Minnesota. Reprinted in Block, N. (ed.). 1980. *Readings in the philosophy of psychology*, vol. 1. Cambridge, MA: Harvard University Press.

Block, N. (ed.). 1980. *Readings in the philosophy of psychology*, vol. 1. Cambridge, MA: Harvard University Press.

Carnap, R. 1928/1961/2003. *Der Logische Aufbau der Welt.* Berlin: Weltkreis-Verlag. Trans. *The logical structure of the world* in *the logical structure of the world and pseudoproblems in philosophy*. Chicago: Open Court.

Carnap, R. 1932a/1934. Die physikalische Sprache als Universalsprache der Wissenschaft. *Erkenntnis* 2: 432–465. Trans. *Unity of science*, London: Kegan Paul.

Carnap, R. 1932b/1959. Psychologie in physikalischer Sprache. *Erkenntnis* 3: 102–142. Trans. Psychology in physical language". In *Logical positivism*, ed. A.J. Ayer, 165–198. New York: Glencoe.

Carnap, R. 1934/2002. *Logische Syntax der Sprache*. Vienna: Springer. Trans. *The logical syntax of language*. Chicago: Open Court.

Carnap, R. 1935/1963. *Philosophy and logical syntax*. London: Routledge and Kegan Paul. Reprinted with addenda added by Carnap and terminological improvements suggested by him, in Alston, W., and G. Nakhnikian (eds.). 1963. *Readings in twentieth century philosophy*, 424–460. New York: The Free Press of Glencoe.

Carnap, R. 1936–1937. Testability and meaning. *Philosophy of Science* 3: 419–471 and 4: 1–40.

Carnap, R. 1938/1991. Logical foundations of the unity of science. In *Encyclopedia and unified science*, ed. O. Neurath et al., 42–62. Chicago: University of Chicago Press. Reprinted in Boyd, Richard, Philip Gasper, and J.D. Trout (eds.). 1991. *The philosophy of science*, 393–404. Cambridge, MA: The MIT Press.

Carnap, R. 1947. *Meaning and necessity*. Chicago: University of Chicago Press.

Carnap, R. 1952. Meaning postulates. *Philosophical Studies* 3: 65–73.

Carnap, R. 1955. Meaning and synonymy in natural languages. *Philosophical Studies* 6: 33–47.

Carnap, R. 1956. The methodological character of theoretical concepts. In *The foundations of science and the concepts of psychology and psychoanalysis*, ed. H. Feigl and M. Scriven, 38–76. Minneapolis: University of Minnesota Press.

Carnap, R. 1963. Nelson Goodman on *Der logische Aufbau der Welt*. In ed. *The philosophy of Rudolf Carnap*, ed. P.A. Schilpp, 944–947. La Salle: Open Court.

Carnap, R. 1964/1994. In *An introduction to the philosophy of science*, ed. Martin Garder. New York: Dover.

Chalmers, D. 2012. *Constructing the world*. Oxford: Oxford University Press.

Chisholm, R. 1955–1956. Sentences about believing. *Proceedings of the Aristotelian Society* 56: 125–148.

Chisholm, R. 1957. *Perceiving. A philosophical study*. Ithaca: Cornell University Press.

Cirera, R. 1993. Carnap's philosophy of mind. *Studies in the History and Philosophy of Science* 24: 351–358.

Cornman, J. 1971. *Materialism and sensations*. New Haven/London: Yale University Press.

Crawford, S. 2013. The myth of logical behaviourism and the origins of the identity theory. In *The Oxford handbook of the history of analytic philosophy*, ed. M. Beaney. Oxford: Oxford University Press.

Creath, R. (ed.). 1992. *Dear Carnap, Dear Van: The Quine-Carnap correspondence and related work*. Berkeley: University of California Press.

Ducasse, C. 1941. *Philosophy as a science. Its matter and its method*. New York: Oskar-Piest.

Feigl, H. 1950/1953. The mind-body problem in the development of logical empiricism. *Revue de Internationale de Philosophie* 4. Reprinted in Feigl, H., and M. Brodbeck (eds.). *Readings in the philosophy of science*, 612–616. New York: Appleton-Century Crofts.

Feigl, H. 1958. The 'mental' and the 'physical'. In *Concepts, theories and the mind-body problem*, ed. H. Feigl, M. Scriven, and G. Maxwell, 370–497. Minneapolis: University of Minnesota Press.

Feigl, H. 1963. Physicalism, unity of science and the foundations of psychology. In *The philosophy of Rudolf Carnap*, ed. P.A. Schilpp, 227–268. La Salle: Open Court.
Feigl, H. 1971. Some crucial issues of mind-body monism. *Synthese* 22: 295–312.
Feigl, H., and W. Sellars (eds.). 1949. *Readings in philosophical analysis.* New York: Appleton Century Crofts.
Feigl, H., M. Scriven, and G. Maxwell (eds.). 1958. *Concepts, theories and the mind-body problem*. Minneapolis: University of Minnesota Press.
Fodor, J. 1968. *Psychological explanation*. New York: Random House.
Goodman, N. 1963. The significance of *Der Logische Aufbau Der Welt*. In *The philosophy of Rudolf Carnap*, ed. P.A. Schilpp, 545–558. La Salle: Open Court.
Hempel, C.G. 1935/1972. Analyse logique de la psychologie. *Revue de Synthese* 10: 27–42. Trans. "The logical analysis of psychology" in Feigl, H., and W. Sellars (eds.). 1949. *Readings in philosophical analysis*, 373–384. New York: Appleton Century Crofts. Reprinted with prefatory note in Marras, A. (ed.). 1972. *Intentionality, mind, and language*, 115–131. Chicago: University of Illinois Press and Block, N. (ed.). 1980. *Readings in the philosophy of psychology*, vol. 1, 14–23. Cambridge, MA: Harvard University Press.
Hempel, C.G. 1951. The concept of cognitive significance: A reconsideration. *Proceedings of the American Academy of Arts and Sciences* 80: 61–77.
Hempel, C.G. 1952. *Fundamentals of concept formation in empirical science*. Chicago: University of Chicago Press.
Hempel, C.G. 1958. The theoretician's dilemma. In *Concepts, theories and the mind-body problem*, ed. H. Feigl, M. Scriven, and G. Maxwell. Minneapolis: University of Minnesota Press. Reprinted in Hempel, C.G. 1965. *Aspects of scientific explanation and other essays in the philosophy of science*, 173–228. New York: The Free Press.
Hempel, C.G. 1965. *Aspects of scientific explanation and other essays in the philosophy of science*. New York: The Free Press.
Hempel C.G. 1969. Logical positivism and the social sciences. In *The legacy of logical positivism*, ed. P. Achinstein and S.F. Barker, 163–194. Baltimore: The John Hopkins Press.
Kim, J. 1971. Materialism and the criteria of the mental. *Synthese* 22: 323–345.
Kim, J. 2003. Logical positivism and the mind-body problem. In *Logical empiricism. Historical and contemporary perspectives*, ed. P. Parrini, W. Salmon, and M. Salmon, 263–280. Pittsburgh: University of Pittsburgh Press.
Leitgeb, H. 2011. New life for Carnap's *Aufbau*? *Synthese* 180: 265–299.
Marras, A. (ed.). 1972. *Intentionality, mind, and language*. Chicago: University of Illinois Press.
Pap, A. 1952. Semantic analysis and psycho-physical dualism. *Mind* 61: 209–221.
Pap, A. 1958. *Semantics and necessary truth*. New Haven: Yale University Press.
Putnam, H. 1963. Brains and behaviour. In *Analytical philosophy, second series*, ed. R. Butler, 211–235. Oxford: Blackwell.
Putnam, H. 1969. Logical positivism and the philosophy of mind. In *The legacy of logical positivism*, ed. P. Achinstein and S.F. Barker, 211–227. Baltimore: The John Hopkins Press.
Quine, W.V.O. 1951. Two dogmas of empiricism. *The Philosophical Review* 60: 20–43.
Quine, W.V.O. 1963. Carnap on logical truth. In *The philosophy of Rudolf Carnap*, ed. P.A. Schilpp, 385–406. La Salle: Open Court.
Schilpp, P.A. (ed.). 1963. *The philosophy of Rudolf Carnap*. La Salle: Open Court.
Searle, J. 2004. *Mind: A brief introduction*. Oxford: Oxford University Press.
Smith, L.D. 1986. *Behaviourism and logical positivism. A reassessment of their alliance*. Stanford: Stanford University Press.
Suppe, F. 1977. The search for philosophic understanding of scientific theories. In *The structure of scientific theories*, 2nd ed, ed. F. Suppe, 3–241. Urbana: University of Illinois Press.

Epistemology Historicized: The French Tradition

Anastasios Brenner

1 Introduction

According to an influential view, the aim of philosophy of science is to provide an analysis of science as encapsulated in theories by means of formal methods. Ideally, theories are axiomatic systems whose empirical content is provided by a definite set of correspondence rules. As stated, this view has of course met with criticism: one should be attentive to the principles guiding scientists in their choice of hypotheses; furthermore, theories should be embedded in larger structures (global theories, paradigms or research programs). This means turning to history of science and taking into account scientific practice. Uncertainty remains however as to the degree of revision required with respect to the methods of philosophy of science as well as its agenda.

In response to this situation a number of scholars are currently promoting what is called "historical epistemology". One can mention Ian Hacking, Lorraine Daston, Hans-Jörg Rheinberger among others. Now, the term historical epistemology employed by these scholars is explicitly borrowed from a group of earlier French thinkers who had formulated a philosophy of science based on history of science. It was coined some 40 years ago in reference to Gaston Bachelard, Georges Canguilhem and Michel Foucault. My aim is to examine how this approach came into existence, what its purpose was and how this French tradition is related to current developments taking place in Berlin, Toronto, New York or elsewhere.

These questions are meant as steps leading to a more fundamental issue: what we may hope to gain from the historical study of science, or in other words the role history should play in philosophy of science. The expression epistemology

A. Brenner (✉)
Centre de Recherches Interdisciplinaires en Sciences Humaines et Sociales CRISES, University Paul Valéry-Montpellier III, Route de Mende, 34199 Montpellier cedex, France
e-mail: anastasios.brenner@wanadoo.fr

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_50,

historicized in my title is to be understood as alluding to Willard Van Quine's article "Epistemology Naturalized" (Quine 1968). He initiated a reorientation of analytic philosophy by denouncing the two dogmas of earlier logical empiricism, namely atomicity and the analytic-synthetic dichotomy. Following thereon post-positivists, such as Thomas Kuhn, Imre Lakatos and Paul Feyerabend, strove to show that the standard view of theories as pure axiomatic systems did not fit the facts. My proposal could be viewed, in a sense, as pushing naturalization further, to include the scientific past. I come then to concur with Rheinberger, who in a recent book, *On Historicizing Epistemology*, calls for a convergence: the historicization of the philosophy of science and the epistemologization of the history of science (Rheinberger 2010, Introduction). I do not propose to provide here and now a study of the French tradition in isolation. My target is the relation of this tradition to mainstream philosophy of science. How can we benefit by taking interest in French philosophy of science, in terms both of positive results and critical insights?

The philosopher is confronted today with a number of challenges. Scientific knowledge has increased greatly, whole new fields of research have come into existence, and novel methods have been set forth. Ensuing technologies have multiplied our means of intervention; they touch more and more directly upon ourselves, upon our bodies and minds. It could be that the tools the philosopher has at his disposal are not sufficient for the task, that what we were taught is in need of been updated. Perhaps philosophy of science is going through a crisis.

Historical epistemology is often contrasted with logical analysis. Yet one may acknowledge the fruitfulness of the latter, while arguing that an exclusive focus on so-called rational reconstructions or formal methods may keep us from grasping some important facets of scientific activity. First, science in its entirety cannot be equated with an axiomatic system. At best it is a series of such systems, and these are continually improved on – hence the dynamics of scientific research. This involves at least some history. Furthermore, logic is a formal science, neighboring on mathematics; it may not provide us with all the adequate tools for understanding the complete encyclopedia of our knowledge.

2 The Origins of Historical Epistemology

In order to characterize more precisely historical epistemology, let us turn toward the context of its introduction. The expression appears in 1969 in the title of a book by Dominique Lecourt, *L'épistémologie historique de Gaston Bachelard* (Lecourt 1969). It designates then Bachelard's method, a philosophical reflection on science that draws on history of science. In response to the debates that this expression gave rise to, Lecourt has more recently provided some explanation of his reasons for having coined it and what he meant thereby. In a book published in 2008, he attributes the paternity to Canguilhem, describing the background in the following terms:

> During the 1960s, when Canguilhem's teaching at the Sorbonne was in particular favor among the students, it became usual to present his work as belonging to the "French" sort of historical epistemology. Precise historical studies have now shown that such a tradition exists, that it developed on the margin and, in some instances, in opposition to what is called Anglo-Saxon epistemology marked by the legacy of the Vienna Circle, logical positivism and the philosophy of language. (Lecourt 2008, p. 51, trans. AB.)[1]

As regards history this passage contains several remarks that need to be unpacked: the mood of the 60s, the interpretation of Bachelard's legacy and the criticism of Anglo-American thought. But let us, at present, follow Lecourt in bringing out the claims Canguilhem was making: the theoretical priority of error, the depreciation of intuition and the concept of the object as a perspective of ideas (Canguilhem 1957). The philosophical orientation is rationalist, Platonist and discontinuist; in addition, there is an inclination to emphasize action. Science is conceived as an elaboration, an overcoming of obstacles, a producing of results, a collective enterprise. Connections were possible with Marxism, which strongly influenced the students of those years. Shortly after introducing historical epistemology, Lecourt went on to include not only Canguilhem, but also Foucault within this category (Lecourt 1972). Because of limited space I cannot provide a complete account of this first stage. Let it suffice to recall one of the arguments Canguilhem gave in favor of recourse to the history of science:

> The specifically philosophical reason hinges on the fact that without recourse to philosophy of science (*épistémologie*) a theory of knowledge would be a meditation on emptiness, and that without any relation to history of science a philosophy of science would be an entirely superfluous double of the science of which it claims to speak. (Canguilhem 1968, p. 11)

This passage, taken from the introduction to a volume of historical and philosophical studies dealing mainly with the life sciences, was written in 1968. Three years later, in the context of post-positivism, Lakatos similarly denounced this separation of philosophy of science and history of science, in his well-known dictum: "Philosophy of science without history of science is empty; history of science without philosophy is blind" (Lakatos 1971, p. 102). Canguilhem deserves credit as an early proponent of the philosophy of the life sciences (Canguilhem 1943; Delaporte 1994). The Bachelardian School thus drew interest to areas of science generally neglected by logical empiricists.[2] Foucault offered, in turn, a philosophy of the social sciences.

As originally developed in the context of French thought, historical epistemology was opposed to logical empiricism. This is not the only possible option. Hacking suggests that history can be conceived as pursuing conceptual analysis by other means. It is clear that the authors who have adopted historical epistemology more recently are not merely seeking to reproduce the work of those forerunners I have

[1] Among the historical studies Lecourt mentions (Bitbol and Gayon 2006), (Brenner 2003).

[2] There are exceptions to this generalization as well as differences between the doctrines of the Vienna Circle and the later versions of logical empiricism it inspired; see, for example (Hofer 2013).

mentioned. Rather they have found here inspiration for opening up new paths, for providing answers to persistent difficulties in philosophy of science today. My intention is to move on to a definition of historical epistemology that is both comprehensive and fruitful.

3 The Reception of Historical Epistemology

We have noted a convergence between historical epistemology and post-positivism. This convergence was however not fully acknowledged at the time. These two currents of thought developed for the most part independently. Canguilhem eventually came across Kuhn's *Structure of Scientific Revolutions* (Kuhn 1962). But his brief and sporadic remarks are highly critical, which will come as a surprise to today's readers (Canguilhem 1977). He was more struck by the differences in their historical approach than their agreement to have recourse to the history of science. Canguilhem appears to have been impatient with Kuhn's reasoning: Bachelard had opened the way for historical philosophy of science some 30 years earlier and it was needless to discuss the theses of the logical empiricists, which were, in his opinion, based on an ill-chosen philosophical option.[3] In turn, Kuhn called exclusively on earlier French thinkers such as Pierre Duhem, Léon Brunschvicg and Émile Meyerson, while ostensibly avoiding any mention of later figures with the exception of Alexandre Koyré (Kuhn 1977). It was only belatedly that Anglo-American authors began to take an interest in the contemporary productions of the French tradition. Browsing through the works of Lakatos, Feyerabend and Hilary Putnam, I found hardly any references to the French authors we have been examining. The exception is a passage in which Putnam mentions the claim that some scientific theories cannot be overthrown by experiment and observations alone, a view he had been defending: "The view is also anticipated by Hanson, but it reaches its sharpest expression in the writings of Thomas Kuhn and Louis Althusser"[4] (Putnam 1975, p. 259).

Kuhn mentions Foucault merely in passing in a late article "What are Scientific Revolutions", in connection with Carl Hempel, Joseph Sneed and Wolfgang Stegmüller: "The resulting picture of linguistic strata shows intriguing parallels to the one discussed by Foucault in *The Archeology of Knowledge*" (Kuhn 1987/2000, p. 14). His most significant statement is in response to the query of a journalist for an interview in the French daily paper *Le Monde*:

[3] Although Kuhn and post-positivists often have only cursory and polemic observations to offer on logical empiricism, there are some attempts at a discussion; see (Suppe 1977).

[4] Althusser was primarily a political philosopher, but some of his texts relate to philosophy of science and contributed to the development of historical epistemology. Teaching for some 40 years, until 1980, at the École normale supérieure, he exerted a strong influence on a whole generation of students. See (Althusser 1974).

> Reading Foucault I came to the conclusion that we are not as proximate as some have suggested. We share indeed the idea that a worldview is determined by a language, which we acquired from Koyré. But there are many differences between my concept of 'paradigm' – a word I have in fact ceased to use, not wishing to endorse all the interpretations that were given to it – and Foucault's notion of episteme. For instance, Foucault never explains how one passes from one episteme to another. (Kuhn 1995, p. 13, trans. AB)

I shall leave the responsibility of this remark to Kuhn, noting that Foucault has a whole section in his *Archeology of Knowledge* devoted to "Change and Transformations" (Foucault 1969, part IV, Chap. 5). Be that as it may, we can say that Anglo-American post-positivists and French historical epistemologists did not really hit it off.

French historical epistemology then did not have an immediate impact abroad. This movement yielded several works that were viewed with some caution at first. After Foucault's early death in 1984 this school lost its momentum. It has practically disappeared in France, to be revived elsewhere. Of course there have been offshoots, later developments relevant to historicizing epistemology.[5] But this is another story.

To be sure the initial program encountered difficulties. French philosophers of science failed to bring their historical method to bear on the debates taking place on the international scene. But more profoundly one perceives in their works a polemic attitude toward formal analysis, which need not be taken up: we can grant that logic is a fruitful tool for philosophy, without sacrificing history; it has become part of the curriculum in France as elsewhere. Why oppose logic and history? One may logically reconstruct the reasoning behind a historical explanation; conversely, one may submit logic to a historical study.[6]

A revival took place. Ian Hacking in his *Historical Ontology* has some instructive remarks to give on the topic (Hacking 2002, p. 9). He tells us that he hit on the term historical epistemology in the 1990s, in seeking to question some received claims in analytic philosophy, only to learn that it had already been employed. The term came to express a new sensitivity shared by several scholars such as Daston, Mary Poovey and Arnold Davidson (Daston 1991; Poovey 1998; Davidson 2001). Foucault was read with enthusiasm, and there was an effort to go back to Canguilhem and even Bachelard.

But it must be understood that the scholars belonging to this group have their own concerns. They are engaged in debates over precise issues with analytic philosophers of science. Hacking brings out the difference with respect to Daston and her followers:

> They study epistemological concepts as objects that evolve and mutate. Their work would be more truly named were it called 'historical meta-epistemology'. Where Bachelard insisted that historical considerations are essential for the practice of epistemology, the historical meta-epistemologist examines the trajectories of the objects that play certain

[5]For an overview of some recent trends – philosophical semantics, rhetorics of science, historical ontology – see (Brenner 2009 and 2011).

[6]Lakatos (1971) offers an example of the former, and we have several histories of logic that give full importance to factors such as the context of elaboration and the chronological development.

> roles in thinking about knowledge and belief [...]. Historical meta-epistemology, thus understood, falls under the generalized concept of historical ontology[7] that I am now developing. (Hacking 2002, p. 9)

The protagonists of this view are more interested in following the evolutions and mutations of concepts then in setting out the history of successive episteme or vast systems of knowledge. One could also turn around Hacking's expression and speak of a *meta*-historical epistemology, that is a second stage of this movement, quite different from the first.

What has happened then in the past 10 or 15 years? This new historical school has produced a large number of works; it has attracted a good deal of attention. Conferences have been staged, bringing together scholars from different parts of the world and different backgrounds. The Max Planck Institute for the History of Science in Berlin has become a prominent center in this respect, having launched a series of conferences on the historical epistemology of the sciences as early as 1995. In a recent study, *Histories of Scientific Observation*, Daston and Lunbeck present their project in the following manner:

> Observation is the most pervasive and fundamental practice of all the modern sciences, both natural and human [...]. Yet scientific observation lacks its own history: why? Countless studies in the history of philosophy of science treat one or another aspect of observation: observation through telescope and microscope, observation in the field or in the laboratory, observation versus experiment, theory-laden observation. But observation itself is rarely the focus of attention and almost never as an object of historical inquiry in its own right (Daston and Lunbeck 2011, p. 1).

The editors refer to Foucault among their sources of inspiration. They direct their criticism at the logical empiricists, adding that post-positivists did not go far enough: historical inquiry as pursued here aims to question the dichotomy between theory and observation from a different angle. Not only are facts to be seen as theory-laden, but fact collecting, observing and experimenting are the result of a practice or activity that has come to acquire its current form through a drawn-out historical process.

The sphere of historical epistemology is expanding more and more. Philipp Kitcher, for example, appears willing to join company (Kitcher 2011). But it would be good to specify his evolution of thought: well versed in analytic philosophy, specializing in philosophy of mathematics, he seeks in his more recent books to open up philosophy of science to ethical and political issues.[8] So historical epistemology has come to designate a large array of research, tending to include whatever departs from a strictly formal philosophy. But one may ask whether there is a genuine historical inquiry here?

Before going on to voice my worries, let me linger a bit on this fashionable trend. We have learned that our scientific worldview is the result of a long, arduous

[7] Hacking borrows this expression from Foucault to designate a study of "the ways in which the possibilities for being arise in history" (Hacking 2002, p. 23).

[8] See (Kitcher 2001, p. XI. Cf. p. 91); significantly, he refers to Foucault (*id.*, p. 53).

and sometimes sinuous path. It is difficult then to employ the ordinary concepts of philosophy of science without stopping to think of their history. We start placing quotation marks around these concepts. They no longer have an obvious meaning and accepted use. They have become problematic.

It could be that, in a broad sense, this research has started to infuse the works of those who do not explicitly acknowledge their debt. If we take for example a late work of a stolid analytic philosopher such as Bas Van Fraassen, *Scientific Representation*, we note that he does not hesitate to recall the debates of the end of the nineteenth century, in order to clarify the concepts he is directing toward current scientific issues, in particular in quantum theory. This return to philosopher-scientists such as Hertz, Maxwell and Duhem allows him to formulate an alternative to mainstream views – his particular brand of empiricist structuralism (Van Fraassen 2008, pp. 80, 92). He even goes back to Renaissance painting and perspective as an important factor in the construction of the scientific representation he is examining. One cannot help recognizing that quite a bit of history goes into some recent variants of analytic philosophy of science.

4 Historical Reflexivity

In developing their program, current advocates of historical epistemology draw freely on previous endeavors. This should not keep us from inquiring into the context in which this program was originally formulated. After all, a historical inquiry should be thorough as regards the evidence at our disposal. What should we think of Canguilhem's reading of Bachelard? He was careful to bring out a significant evolution of thought. According to Canguilhem, Bachelard moved away from the idealism that had dominated earlier philosophy of science. The "applied rationalism" and "instructed materialism" of his later thought made it possible to truly take into consideration other aspects of science. His "phenomenotechnics" could be seen as a call for a precise description and careful analysis of the material aspects of science. This led to an interest in machines and technology that his followers were to take up. After highlighting these themes in Bachelard's thought, Canguilhem gave this summary: "Scientific proof is labor because it reorganizes the given, because it provokes effects that have no natural counterpart, because it constructs its sense organs" (Canguilhem 1968, p. 192).

Let us seek to locate this reading in the setting of French philosophy after World War II. I do not believe that Canguilhem was claiming to give a precise rendition of Bachelard; he was pursuing his own philosophical agenda (Canguilhem 1977, p. 9). So, it appears important to go back to Bachelard's original thought. Studying his early texts, we learn that he derived many ideas from Poincaré, Duhem and Abel Rey. He thus shared with the logical empiricists some sources of inspiration. Furthermore, one does find references to logical empiricism in his works. These are sparse but generally favorable. Bachelard was impressed by the interpretations Schlick and Reichenbach offered of Relativity theory, even if he went

on to introduce his own claims (Bachelard 1929, p. 188; 1934, p. 65). It appears that his followers were responsible for giving a polemic twist to his thought.

One should broaden the scope of the inquiry. For the issues we are concerned with involve many other thinkers. We must reach back at least to the debate between Poincaré and Russell. Furthermore, historical epistemology could well be applied to the works of Bachelard's teachers, Léon Brunschvicg and Abel Rey.[9] What I am suggesting is that we need to do more historical inquiry in order to understand what went on here, for it is a major problem, that of the divide between the analytical approach and the historical approach. And I believe this is what one expects of a historical epistemology that pursues its method to a full degree.[10]

Historical epistemology is an open program. One may add on to the various histories of objects, concepts and practices, with a view to drawing up a new encyclopedia, each entry comprising its historical section. More in-depth studies are indeed called for. Daston and Galison have written a well-informed and significant history of the concept of objectivity. After reading the book, one may still ask what lies behind their choice of topic, what makes their study relevant for us. For sure, contemporary science aims at objectivity. But this is only one among several rational values or epistemic virtues. Objectivity is no different from other constitutive values such as accuracy, consistency, simplicity or predictive power. One should then submit all these different values to a careful scrutiny. They are the result of a historical process. Now, these notions are not unrelated to the development of the axiomatic outlook. Indeed, the axioms or postulates of a theory are freely chosen. They would be arbitrary if we did not have values to guide us. Daston and Galison's study is thus related, albeit polemically, to this conception. But they do not bring out clearly enough the connection. My point is that we need a self-conscious and explicit history of philosophy of science as a means of locating our position, that is the perspective from which our inquiry is carried out.

5 Conclusion

Should my demand for historical detail appear punctilious, let me evoke a few concrete concerns. Having worked with practicing scientists, I know that a philosopher may have difficulty in responding to their requests. Scientists tend to be surprised if not bored by finessing on observation sentences. They are reluctant, when philosophizing, to add an extra layer of formalization with mathematical logic. They tend to be more interested in history, and often even the pre-paradigmatic history of their discipline. Here they find intriguing metaphysical problems, unfamiliar ontologies. This is not only more enjoyable for them, but also more useful in their

[9] For a recent study of Brunschvicg, see (Chimisso 2008).

[10] One can draw here on various studies devoted to the history of logical empiricism on the one hand and French philosophy of science on the other. See (Uebel 2003), (Brenner 2003).

endeavor to go beyond received solutions and current frameworks. By concentrating on the context of justification, philosophy of science has led us away from the quintessence of scientific activity, that is discovery, invention and innovation.

Furthermore, philosophy is taught in secondary education in France as well as in Italy and other European countries. Our job as university professors also includes training students who will later be teaching at this level. Philosophy of science belongs to this curriculum. And we would surely be making our task difficult if we were to restrict ourselves to a strictly logical analysis directed to the structure of axiomatic systems. To give an enlightening picture of science and its consequences – methodological, ethical and political – we need to approach it from many directions. Bachelard and Canguilhem had taught at *lycées* before holding university chairs. They were very attentive to the issue of bringing a young audience to an understanding of their specialty.[11] Their qualms over a formal philosophy of science excluding historical comprehension are undoubtedly rooted in this preoccupation, and their criticism deserves to be recalled. We should then conceive the method of philosophy of science with this in mind. We also encounter the question of bringing into contact the different philosophical traditions. This is a central problem in the context of the European union as a democratic entity. How can we formulate a critical reflection on science and technology, a free inquiry into these activities, which are deeply changing our everyday lives? As a means of clearing up misunderstandings and developing a broad scope, I believe that historical epistemology or epistemology historicized remains an essential tool.

References

Althusser, L. 1974. *Philosophie et philosophie spontanée des savants*. Paris: Maspero. Trans. W. Montag *Philosophy and the spontaneous philosophy of the scientists*. London: Verso, 1990.

Bachelard, G. 1929. *La valeur inductive de la relativité*. Paris: Vrin.

Bachelard, G. 1934. *Le nouvel esprit scientifique*. Paris: PUF, 1971. Trans. A. Goldhammer *The new scientific spirit*. Boston: Beacon Press, 1984.

Bachelard, G. 1938. *La formation de l'esprit scientifique*. Paris: Vrin. 1975.

Bitbol, M., and J. Gayon (eds.). 2006. *L'épistémologie française*: 1930–1970. Paris: Presses universitaires de France.

Brenner, A. 2003. *Les origines françaises de la philosophie des sciences*. Paris: Presses universitaires de France.

Brenner, A. 2009. A problem in general philosophy of science: The rational criteria of choice. In *French studies in the philosophy of science*, ed. A. Brenner and J. Gayon, 73–90. Vienna: Springer.

Brenner, A. 2011. *Raison scientifique et valeurs humaines*. Paris: Presses universitaires de France.

Canguilhem, G. 1943. *Le normal et le pathologique.* Paris: Presses universitaires de France, 2005. Trans. C.R. Fawcett *The normal and the pathological*. New York: Zone Books, 1989.

[11] See for example (Bachelard 1938, ch. 12). Canguilhem wrote an international survey on the teaching of philosophy: (Canguilhem et al. 1953).

Canguilhem, G. 1957. Sur une épistémologie concordataire. In *Hommage à Gaston Bachelard: Études de philosophie et d'histoire des sciences*, ed. C. Bouligand et al., 3–12. Paris: Presses universitaires de France.
Canguilhem, G. 1968. *Études d'histoire et de philosophie des sciences*. Paris: Vrin.
Canguilhem, G. 1977. *Idéologie et rationalité*. Paris: Vrin.
Canguilhem, G., et al. 1953. *L'enseignement de la philosophie: Enquête international*. Paris: UNESCO.
Chimisso, C. 2008. *Writing the history of the mind: Philosophy and science in France 1900 to 1960s*. Aldershot: Ashgate.
Daston, L. 1991. Objectivity and the escape from perspective. *Social Studies of Science* 22: 597–618.
Daston, L., and E. Lunbeck (eds.). 2011. *Histories of scientific observation*. Chicago: Chicago University Press.
Davidson, A. 2001. *The emergence of sexuality: Historical epistemology and the formation of concepts*. Cambridge, MA: Harvard University Press.
Delaporte, F. (ed.). 1994. *A vital rationalist: Selected writings from Georges Canguilhem*. New York: Zone Books.
Foucault, M. 1969. *L'archéologie du savoir*. Paris: Gallimard. Trans. A.M. Scheridan Smith, New York: Pantheon, 1971.
Hacking, I. 2002. *Historical ontology*. Cambridge, MA: Harvard University Press.
Hofer, V. 2013. Philosophy of biology in early logical empiricism. In *New challenges to philosophy of science*, ed. H. Andersen et al. Dordrecht: Springer.
Kitcher, P. 2001. *Science, truth and democracy*. Oxford: Oxford University Press.
Kitcher, P. 2011. Epistemology without history is blind. *Erkenntnis* 75: 505–524.
Kuhn, T. 1962. *The structure of scientific revolutions*. Chicago: Chicago University Press. 1970, 2nd ed.
Kuhn, T. 1977. *The essential tension*. Chicago: Chicago University Press.
Kuhn, T. 1987. What are scientific revolutions? In *The probabilistic revolution, vol.1: Ideas in history*, ed. L. Krüger, L. Daston, and M. Heidelberger. Cambridge, MA: The MIT Press. Reprinted in *The road since structure*, 11–32. Chicago: Chicago University Press, 2000.
Kuhn, T. 1995. Un entretien avec Thomas S. Kuhn. Edited and translated by C. Delacampagne, *Le Monde*, 5–6 February: 12–13.
Lakatos, I. 1971. History of science and its rational reconstructions. In *The methodology of scientific research programmes: Philosophical papers I*. Cambridge: Cambridge University Press. 1978.
Lecourt, D. 1969. *L'épistémologie historique de Gaston Bachelard*. Paris: Vrin.
Lecourt, D. 1972. *Pour une critique de l'épistémologie: Bachelard, Canguilhem, Foucault*. Paris: Maspero.
Lecourt, D. 2008. *Georges Canguilhem*. Paris: Presses universitaires de France.
Poovey, M. 1998. *A history of modern fact*. Chicago: Chicago University Press.
Putnam, H. 1975. *Philosophical papers I*. Cambridge: Cambridge University Press.
Quine, W.V.O. 1968. Epistemology naturalized. In *Ontological relativity*, 69–90. New York: Columbia University Press. 1969.
Rheinberger, H.-J. 2010. *On historicizing epistemology* (Trans. D. Fernbach). Stanford: Stanford University Press.
Suppe, F. (ed.). 1977. *The structure of scientific theory*. Chicago: University of Illinois Press.
Uebel, T. 2003. On the Austrian roots of logical empiricism: The case of the first Vienna circle. In *Logical empiricism: Historical and contemporary perspectives*, ed. P. Parrini et al., 67–93. Pittsburgh: University of Pittsburgh Press.
Van Fraassen, B. 2008. *Scientific representation*. Oxford: Oxford University Press.

Commentary on Anastasios Brenner's "Epistemology Historicized"

Cristina Chimisso

1 Introduction

In his paper, Anastasios Brenner questions the way in which we do philosophy of science, and invites us to have a more reflexive approach to the history of philosophy of science. He touches on many crucial problems, ranging from the way we approach the study of science, to the role that philosophy of science – or more precisely historical epistemology – should have in education and indeed society. I shall only comment on a couple of points, which seem important to me. The first question lies at the very core of historical epistemology, the type of philosophy of science on which Brenner focuses. This is the relation between history and philosophy. As Brenner points out, the French tradition in philosophy of science that has produced "classic" historical epistemology is perhaps the best example of how attention to history changes the way we see science.[1] For this reason, it has been a point of reference for current scholars who also aim to integrate history and philosophy of science. I think that this integration is a worthwhile and indeed necessary goal, but also that it is perhaps even more complex than it seems. I shall mention some examples to show that even in classic historical epistemology this integration was far more problematic and imperfect than perhaps it appears. The second question, related to the first, concerns history of philosophy. Brenner thinks, rightly in my view, that philosophers of science need history of philosophy of science to locate their own work, and that history of philosophy of science

[1] I call "classic" historical epistemology the philosophical movement that developed in France in the twentieth century, and that notably includes Gaston Bachelard and Georges Canguilhem. I call it "classic" in order to distinguish it from current historical epistemology.

C. Chimisso (✉)
Arts Faculty, Department of Philosophy, The Open University, MK7 6AA Milton Keynes, UK
e-mail: cristina.chimisso@open.ac.uk

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_51,

plays an important pedagogical and political role. But what type of history of philosophy should we practice? History of philosophy includes diverse approaches; what history means in many of them is rather different from what it means for general historians, among others. Current historical epistemology has as its object history of science, and refers to past philosophies in order to interpret this object. In other words, it refers to the past of science and to the past of philosophy, although those occupy different places within it. I shall make some remarks on the different ways in which the past of science and the past of philosophy are received, that is to say on the difference between history of science and history of philosophy. I shall argue that a variety of approaches in the study of philosophy and science is welcome, and that the use we intend to make of the past influences the way we approach it. However, I shall also argue that for history of philosophy to play the role that Brenner envisages, it should be genuinely historical, unlike most of its mainstream versions, and less dissimilar from history of science. I shall conclude by raising some problems with the historical approach that I advocate, and by briefly proposing some possible ways in which we may be able to address them.

2 The Difficult Relation of History and Philosophy

Brenner invites us to write a history of the French tradition and to retrace the debate back from Canguilhem to Bachelard, Poincaré, Brunschvicg and Rey. On this, he is certain to find me in agreement, as, like him, I have dedicated much work to this history. The central feature of large part of French philosophy of science, the interaction of history and philosophy, is also its most appealing, to me and to many others. The particular institutional, educational and indeed social history of the first half of the twentieth century in France brought history and philosophy close in a way that has not been the case in other traditions.[2] Classic historical epistemology originated and developed in a favourable environment. Institutionally, it found a fertile soil as a range of disciplines, including ethnology, sociology, psychology and history of science were in the process of finding their own identity while maintaining a close relationship with philosophy. It goes without saying that the close relationship of philosophy with history of science is the most apparent and fundamental. The emergence of history of science as an autonomous discipline with its institutes, journals, jobs and qualifications was securely linked to philosophy. The directors of the Institute of History of Science and Technology at the University of Paris and of its journal, *Thalés*, were philosophers, and included Abel Rey, Gaston Bachelard and Georges Canguilhem.

From an intellectual point of view, what made philosophy interact not only with history of science but also with ethnology, sociology and psychology was that all these disciplines addressed similar questions. The investigation of how the human

[2] I discuss the history and development of this tradition in (Chimisso 2008).

mind works was at the very core of the work of a number of these disciplines' practitioners. The empirical data on the functioning of the mind could be gathered in different ways: psychologists set up laboratories, the historian of philosophy Lucien Lévy-Bruhl turned to ethnological reports (and co-founded the Institute of Ethnology at the University of Paris) while other philosophers turned to the history of science.[3] The close interaction of philosophy and history developed in this interdisciplinary milieu. Moreover, the use of a historical model in philosophy was legitimised by a very French tradition, that of Positivism, and in particular by Auguste Comte's conception of the mind as having a history.

However, the history of philosophy provided further legitimizing models for the interaction of history and philosophy of science. Indeed, Georges Canguilhem argued that philosophy of science had always paid close attention to the history of science. But for him epistemologists did not always grasp that their exemplar was not the definitive expression of science, but rather a particular historical episode. He cited Kant who, in the second Preface of the *Critique of Pure Reason*, had justified his critical project by reference to the history of mathematics and physics. However, Canguilhem regarded Kant as fundamentally mistaken in his belief that he could extract a definitive set of norms governing the production of knowledge from the science of his time (Canguilhem 1993 [1977], pp. 19–20). On this point, as in many others, Canguilhem invoked Bachelard's authority, but did not cite Léon Brunschvicg who in fact had made the very same point. As a self-proclaimed neo-Kantian, Brunschvicg aimed to study the mind as Kant had done, but with a crucial difference: for him no single episode of the history of science could give us the key to understanding how the mind works. Rather, an extensive study of the history of science should serve this purpose. The history of science for him showed that, contrary to what Kant had concluded, the mind changes in time. For the philosopher who believed that the concepts that structure our knowledge are not given once and for all, epistemology could not be done without history (Brunschvicg 1922, p. 552; Brunschvicg 1936). Indeed, even those who did not think that the mind changes in significant ways, as for instance Emile Meyerson, employed history in order to prove their point (Meyerson 1931). The question of whether the mind changes according to time and culture was on the table, and history was necessary in order to answer it.

The interaction of history and philosophy in the French tradition has created a particular perspective not only on science, but on epistemology, in the English rather than French meaning, that is on the theory of knowledge in the broad sense. The legacy of this tradition has taken different forms, and notably nowadays it is invoked by the scholars working within the modern version of historical epistemology (Rheinberger 2005, 2010b; Hacking 1999, 2002; Davidson 1998; Daston 2000). I strongly believe in the potential of the marriage of philosophy and history, and

[3]Here are some examples of the use of ethnological reports and of the history of science in the study of the mind: (Lévy-Bruhl 1910, 1922, 1996 [1927]); (Brunschvicg 1912, 1922); (Bachelard 1991 [1934], 1993 [1938], 1988 [1940]); (Metzger1926); (Canguilhem 1943, 1955).

also in the teaching and inspiration with which the French tradition can provide us in this context. However, this marriage is a difficult one, and we should not idealize the occasions in which it seems to have worked. It was difficult in the French tradition as well, and we cannot learn from it without a critical approach. In fact, some of the issues that Brenner indicates regarding the present situation can be easily found there too. The classic historical epistemologists' advantage was perhaps that those among them more committed to history and those more committed to philosophy were in close contact, shared students and seminars sessions.[4] This closeness, however, also brought to the fore differences, although these were not always easily expressed, as the balance of power and the web of dependency or friendships were often difficult to ignore.

Hélène Metzger is a case in point. She was in a rather weak academic position for all her life, despite the high quality of her publications and her commitment to the teaching of the students of the Institute of History of Science and Technology of the University of Paris where historical epistemology developed. As a historian of chemistry, she not only chose philosophers as her interlocutors, but also produced important epistemological works (Metzger 1926, 1987). She shared with the philosophers theoretical concerns, namely understanding the different ways of thinking in the history of science.[5] Despite her close collaboration with philosophers, and the support she received from them, she occasionally stressed the shortcomings in their use of history. She chose to express her doubts most clearly in a conference paper on the philosopher Emile Meyerson, who regarded himself as her mentor (a role that she in fact contested, see Chimisso and Freudenthal (2003)). She stressed that his work was epistemological (rather than historical), and that his use of "historical examples" aimed at understanding the mind was part of a tradition, that included philosophers as diverse as Comte, Cournot, Renouvier, Mach, Duhem, and "her colleagues" Abel Rey and Léon Brunschvicg. She welcomed the collaboration of history and philosophy, but also emphasised that the history of science must be "scrupulously" and wisely interpreted, and never distorted, in order to be of use to the philosopher (Metzger 1987, pp. 95–106). I do not think that there is much doubt that she thought that Meyerson and other philosophers used historical "examples" in order to confirm their theses (which were rather diverse), but did not have enough concern for historical accuracy and research. Her own intensive work on sources, attention to details and her narrow historical focus, mainly on seventeenth and eighteenth-century history of chemistry, greatly differed from the philosophers' grand narratives.

While Metzger was very much part of the milieu in which historical epistemology developed, and indeed can be seen as a historical epistemologist, the historian of

[4] I analyze the links between the scholars in the historical epistemology milieu and surroundings in (Chimisso 2008), and in (Chimisso 2001). The minutes of the seminars held at the Société française de philosophie, published in its *Bulletin*, are very interesting in this, and others, respects.

[5] Her lectures on Newton at the École Pratique des Hautes Études are arguably the best example of the historical application of her theoretical concerns (Metzger 1938).

mentalités Lucien Febvre, founder of the *Annales*, kept his distances from that milieu, and indeed from philosophers of science in general, despite his close links with Abel Rey. His attack on the philosophers' use of history was particularly caustic, and he dismissed the work of those philosophers, like Brunschvicg, whom he regarded as old-fashioned, and only concerned with "high" culture. He concluded that what those philosophers called history had little do with the history of historians like himself (Febvre 1992 [1948]). Febvre's unwillingness to engage with historical epistemology, which in fact developed from Brunschvicg's teachings, has been seen as a missed chance by the historian Roger Chartier. The latter has argued that if Febvre had engaged with it, he could have avoided theoretical naiveties in his own work (Chartier 1988, pp. 35–36). The fact remains that a sustained dialogue did not take place, either then or later. To mention a more recent example, Michel Foucault's work – which springs from this tradition – has been criticised by historians for his obscure style, his lack of narrative, his historical inaccuracies, his ambiguous relationship with truth, his disregard for causal explanations, and more (Rowlinson and Carter 2002). Indeed, Mark Poster has commented that "many American and British historians have received Foucault's books ... as an attack on the discipline of history" (Poster 1984, p. 73).

Classic historical epistemology was developed mainly by philosophers who worked in close contact with ethnologists, psychologists and sociologists, especially at the beginning of the history of this tradition. Their history was history of science. As a discipline, history of science was developed by the philosophers themselves, and kept rather apart from general historians, including historians of *mentalités*, who arguably should have been their obvious interlocutors. Perhaps it was not only historians like Febvre who missed an opportunity by not engaging with the philosophers, as Chartier argues, but also the philosophers by not learning from the historians. The mutual distrust and lack of knowledge of one another's work meant that historians were more likely to either ignore or attack classic historical epistemology. The current version of historical epistemology, on the other hand, appears to have developed in historical milieus, notably at the Max Planck Institute for History of Science in Berlin, and to be more dominated by historical approaches – although a spectrum of approaches exists in it as it did for classic historical epistemology, as shown by Lorraine Daston's and Ian Hacking's respective works. It could be argued that the space that history and philosophy respectively occupy in modern historical epistemology is the mirror image of that they occupy in classic historical epistemology. This is of course also because although the two schools share the name "historical epistemology" and current historical epistemologists refer to Bachelard and Canguilhem, their aims are different. The aim of classic historical epistemology, and first of all of Bachelard, was to answer philosophical questions about the mind, and to produce a theory of knowledge. Current historical epistemology does not appear to be epistemology in the traditional sense, but rather a history of epistemological objects. Ian Hacking has commented on this difference and has proposed "historical meta-epistemology" for current historical epistemology (Hacking 1999, 2002, pp. 9ff). However, this distinction between the two historical epistemologies may paper over the variety of

approaches within them; things may look more complex if we for instance include Metzger and Foucault within classic historical epistemology.

3 How Historical Is History of Philosophy?

Brenner argues that in order to locate their own approach, philosophers of science need a "self-conscious and explicit history of philosophy of science". But what type of history of philosophy would help us to do so? Recent historical epistemology, as Brenner reminds us, has referred to ideas coming from previous traditions, first of all, but not exclusively, the French tradition that I have discussed above. Many scholars have used these ideas as a framework for their own novel approach to history, drawing interesting conclusions that in turn interest the philosopher. The question however arises: how should we approach past philosophical ideas? Is their origin meaningful, or do they just provide timeless answers to timeless questions? Is history of philosophy a dialogue with great ideas that have been formulated in the past, but that are ultimately current? Or perhaps past ideas are not timeless, but we can just appropriate them and adapt them to our own context? Many philosophers regard their own discipline as not having a history in the sense of either development or change; for them history of philosophy is rather a repository of texts and ideas with which we engage in dialogue, regardless of their origins. There are various degrees of this approach, including an almost completely ahistorical one. Many of us use this approach in teaching: for instance, David Hume is often introduced to students in order to teach the issue of the justification of induction, without a particular historical perspective. "Hume" becomes a short-hand for what is regarded as a long-lasting philosophical problem. Much of history of philosophy pays more attention to history than that, but most of the time in a very restricted way. Often it is either analysis of texts, or what historians of science would call internal history. If related to something outside themselves, philosophical ideas are related to previous, contemporary and current philosophical ideas, as if they could only generate one another. In Ian Hacking's words, "history of philosophy practised in universities [is] committed to philosophical epochs and schools, and dedicated to a canonical list of philosophers whom it regards as pen pals across the centuries" (Hacking 2002, p. 6). Here for the sake of brevity and clarity I am presenting a rather simplified image of some of the approaches existing within history of philosophy, but my image is no straw man.

Historical epistemology itself is not unaffected by this problem. Brenner cites Hans-Jörg Rheinberger's excellent book, *On Historicizing Epistemology* (Rheinberger 2010b). In particular, Brenner agrees with Rheinberger when the latter calls for a convergence between history and philosophy of science; this in his view should be achieved by historicizing philosophy of science and by epistemologizing the history of science. This is an ambitious and very attractive programme, and its two sides appear to include the various projects of classic and current historical epistemology. However, inevitably there have been and there will be different ways

of intending both the epistemologization of history of science and the historicization of philosophy of science. How do we historicise philosophy of science? Can the history of philosophy of science be approached as we do the history of science? I am bracketing here all the difficulties and anachronism of boundaries between the disciplines that now we label science and philosophy. The current disciplines of history of science and history of philosophy have distinct, if at times overlapping, objects. The historicity of these objects does not appear to be taken equally seriously. Even Rheinberger seems to approach philosophy and science rather differently. To mention just an example: when writing about scientific instruments, Rheinberger emphasizes the role of the "historical and local" context in which they are embedded (Rheinberger 2010a). Needless to say, Rheinberger does not construct his history of epistemic things as a linear and progressive narrative. On the other hand, when confronted with the history of philosophy, he adopts a rather different perspective. In *On Historicizing Epistemology* he offers a very insightful treatment of a number of philosophers who have inspired current historical epistemology. In this case, however, there is no context for the emergence of ideas, which are arranged in a rather linear development, and the scholars discussed seem to share ideas and approaches across places and traditions.[6] Even in the hands of such an eminent representative of historical epistemology, the history of philosophy appears as a repository of ideas, or as having at least a rather strict internal history.

However, it is not simply the case that Rheinberger somewhat falls back on more traditional ways when approaching philosophy. His different approaches reveal a very crucial problem that is not easy to solve: to which extent can we historicize philosophical ideas? If we approach past (or present) philosophical ideas in order to inform our work, for instance in order to write a history (of science, or indeed philosophy), how can we historicize those very ideas that are supposed to guide us in the way in which we approach the object of our study? In other words, when historians of science rely on historiographical and epistemological ideas in order to make sense of past documents and objects, how can they make these very ideas objects of their study as well? This is not only a problem for the historian, but also for the philosopher; in fact, Brenner's comments are mainly about the philosopher of science's work. Indeed, philosophers of science have two sets of problems, as they deal both with scientific and philosophical ideas. Classic historical epistemologists accepted the historicity of science and constructed their epistemologies accordingly. Newtonianism could no longer be the only model of knowledge for them: the daring changes that were occurring in science, notably non-Euclidean geometries, the theory of relativity and quantum mechanics, indicated to them that the history of science was not a cumulative progress, and more importantly, that the organising concepts of human knowledge were not stable. Indeed, they historicised previous philosophical attitudes towards science: Canguilhem, in the passage cited above, briefly explained Kant's view of the immutability of intellectual categories as a cultural product of the Enlightenment. For him "it would have been difficult" at the

[6] I discussed this in my review of Rheinberger's book (Chimisso 2012).

time to entertain the possibility of a history of the categories of scientific thought (Canguilhem 1993 [1977], p. 20). But what about Canguilhem's own view? When we write a history of philosophy of science, the view that knowledge is historical should also be treated as historical, as emerging thanks to particular intellectual, historical and institutional circumstances. In turn, the assumptions of the historian of philosophy of science are also historical products, and so on. Is it there a risk of an infinite regress?

4 Conclusion

I shall not pretend to have a simple answer, indeed any proper answer to these difficult questions. I only have a couple of points I would like to address. First of all, I wish to clarify that I believe that a variety of approaches to philosophy, and different degrees of historical and/or sociological perspectives should be welcome. After all, the aims of research vary. Just as scientists involved in a particular research cannot at the same time write a history of scientific knowledge, and question the fundamental assumptions of their work, so philosophers may have to bracket the historicity of their concepts when dealing with a theoretical issue. At the same time, I also believe that a more genuinely historical approach to philosophical ideas, and above all an awareness of the historicity of our assumptions and methods is needed. Philosophical ideas, just as scientific ideas, emerge at a certain point in history, under specific social, educational and cultural conditions, and develop historically. A history of philosophy that takes history seriously is necessary in order to acquire a specific awareness of the historicity of philosophical ideas, practices and methods. This type of history of philosophy looks rather different from the more traditional history of philosophy, of the type that is found in history of philosophy textbooks. A genuinely historical history of philosophy could learn from science studies that a sociological approach can also help to make sense of the emergence and development of philosophical knowledge. Despite excellent works have been published in the recent years, the sociology of philosophy and the sociology of philosophical knowledge are rather small fields, and interestingly there seems to be a greater resistance to a sociological study of philosophy than of science.[7]

To return to the issue of regress, the type of history of philosophy I am promoting can obviously be subject to it. However, it provides a degree of reflexivity that may enable its practitioner to escape it at least partially. This is because by studying philosophy's historical production, the historian of philosophy is also in a good position to keep a critical eye on her own practices and ideas. She would not only check her own ideas against past ideas, in the timeless dialogue many of us have been trained to perform, but also the conditions of production of her own ideas and

[7] For a bibliography of prominent works in these areas, see (Heidegren and Lundberg 2010).

methods against those she studies. A system of cross-controls, as Pierre Bourdieu has put it with regard to sociology (Bourdieu 2004), can be implemented, and can give a solid foundation to the study of philosophy. In short, I agree with Brenner that a history of philosophy of science would enable philosophers to locate their own work, and, like him, I welcome a "historicized epistemology". But I would like to add that history of philosophy, in order to play this role, should be fully and properly historical. Only a genuinely historicized history of philosophy would provide the philosopher with a powerful reflexive tool.

References

Bachelard, G. 1988 [1940]. *La philosophie du non. Essai d'une philosophie du nouvel esprit scientifique*. Paris: Presses Universitaires de France.

Bachelard, G. 1991 [1934]. *Le nouvel esprit scientifique*. Paris: Presses universitaires de France.

Bachelard, G. 1993 [1938]. *La formation de l'esprit scientifique: contribution à une psychanalyse de la connaissance objective*. Paris: Vrin.

Bourdieu, P. 2004. *Science of science and reflexivity*. Cambridge: Polity.

Brunschvicg, L. 1912. *Les étapes de la philosophie mathématique*. Paris: Alcan.

Brunschvicg, L. 1922. *L'expérience humaine et la causalité physique*. Paris: Alcan.

Brunschvicg, L. 1936. History and philosophy. In *Philosophy and history. Essays presented to Ernst Cassirer*, ed. R. Klibansky, 27–34. Oxford: Clarendon.

Canguilhem, G. 1943. *Essai sur quelques problèmes concernant le normal et la pathologique*. Strasbourg: Publications de la Faculté des Lettres de Strasbourg.

Canguilhem, G. 1955. *La formation du concept de réflexe aux XVII^e^ et XVIII^e^ siècles*. Paris: Presses Universitaires de France.

Canguilhem, G. 1993 [1977]. *Idéologie et rationalité dans l'histoire des sciences de la vie*. Paris: Vrin.

Chartier, R. 1988. *Cultural history. Between practices and representations*. Ithaca: Cornell University Press.

Chimisso, C. 2001. Hélène Metzger: The history of science between the study of mentalities and total history. *Studies in History and Philosophy of Science* 32A(2): 203–241.

Chimisso, C. 2008. *Writing the history of the mind: Philosophy and science in France, 1900 to 1960s*. Aldershot: Ashgate.

Chimisso, C. 2012. What is historical epistemology? *Radical Philosophy* 171: 36–39.

Chimisso, C., and G. Freudenthal. 2003. A mind of her own: Hélène Metzger to Emile Meyerson, 1933. *Isis* 94(3): 477–491.

Daston, L.J. 2000. *Biographies of scientific objects*. Chicago/London: University of Chicago Press.

Davidson, A.I. 1998. *The emergence of sexuality: Historical epistemology and the formation of concepts*. Cambridge, MA: Harvard University Press.

Febvre, L. 1992 [1948]. Un cours de Léon Brunschvicg. In *Combats pour l'histoire*, 289–924. Paris: Colin.

Hacking, I. 1999. Historical meta-epistemology. In *Wahrheit und Geschicte*, ed. W. Carl and L. Daston, 53–77. Göttingen: Vandenhoeck & Ruprecht.

Hacking, I. 2002. *Historical ontology*. Cambridge, MA: Harvard University Press.

Heidegren, C.-G., and H. Lundberg. 2010. Towards a sociology of philosophy. *Acta Sociologica* 53(3): 3–18.

Lévy-Bruhl, L. 1910. *Les fonctions mentales dans les sociétés inférieures*. Paris: Alcan.

Lévy-Bruhl, L. 1922. *La mentalité primitive*. Paris: Alcan.

Lévy-Bruhl, L. 1996 [1927]. *L'âme primitive*. Paris: Presses Universitaires de France.

Metzger, H. 1926. *Les concepts scientifiques*. Paris: Alcan.
Metzger, H. 1938. *Attraction universelle et religion naturelle chez quelques commentateurs anglais de Newton*. Paris: Hermann.
Metzger, H. 1987. *La méthode philosophique en histoire des sciences. Textes 1914–1939, réunis par Gad Freudenthal*. Paris: Fayard.
Meyerson, E. 1931. *Du cheminement de la pensée*. Paris: Alcan.
Poster, M. 1984. *Foucault, Marxism and history: Mode of production versus mode of information*. Cambridge: Polity.
Rheinberger, H.-J. 2005. Gaston Bachelard and the notion of 'phenomenotechnique'. *Perspectives on Science* 13(3): 313–328.
Rheinberger, H.-J. 2010a. *An epistemology of the concrete: Twentieth-century histories of life*. Durham: Duke University Press.
Rheinberger, H.-J. 2010b. *On historicizing epistemology: An essay*. Stanford: Stanford University Press.
Rowlinson, M., and C. Carter. 2002. Foucault and history in organization studies. *Organization* 9(4): 527–547.

History and Philosophy of Science: Between Description and Construction

Friedrich Stadler

1 History and Philosophy of Science (HPS): Institutions and Individuals

A Google search of the terms "History and Philosophy of Science" yields an impressive number of hits indicating the international presence of the research field, as does one for the German words "Wissenschaftsgeschichte und Wissenschaftsphilosophie".[1] A further search for current study programs (Master's and Doctorate programs) in the field reveals about 45 in Europe, 35 in North America, and 4 each in Australia and Israel. As for journals more or less strongly associated with the field, 170 can be located in Europe.[2]

This is surprising given that this hybrid field of teaching and study has not acquired a clear disciplinary identity in terms of its methods and subject range. It is clear, however, that HPS is not simply a merger of history of science and philosophy of science but a meeting of two distinct scientific cultures where a largely positive connotation and institutional presence is accompanied by the theoretical porosity of an inter- and trans-disciplinary field. There is no lack of postulate-like claims about the purpose and productivity of HPS since the historical-sociological turn in

This article is a shortened and revised English version of my "History and Philosophy of Science", in: *Berichte zur Wissenschaftsgeschichte* 35 (2012), pp. 217–238 (Stadler 2012a).

[1] Accessed 19.6.2012: 71,700,000 and 782,000. In the English Wikipedia, they speak casually of an "academic discipline".

[2] ESF, ERIH-Index 2007. In the current index from 2009, one now only finds journals of philosophy of science, because the historians of science left the HPS-panel in sign of protest against the ERIH project, see the editorial in *Berichte zur Wissenschaftsgeschichte* 32, 2 (2009): 131–134.

F. Stadler (✉)
Institute Vienna Circle, University of Vienna, Spitalgasse 2, Hof 1.13, Vienna A-1090, Austria
e-mail: Friedrich.Stadler@univie.ac.at

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1_52,

the theory of science in the 1960s.[3] (Since these issued primarily from philosophers of science, one can ask whether historians of science view the matter in the same way.) The classification adopted by official bodies awarding research grants also reflects a lack of systematicity: for instance, history and philosophy of science are sometimes classed under "humanities", sometimes under "historical sciences".[4] In light of this variability it is legitimate to ask to what extent the institutional placement and scientific location of HPS in history or philosophy represents a non-accidental alignment. At least in the case of Thomas S. Kuhn it became clear that his affiliation which fluctuated between philosophy and history departments was an important factor in defining both his professional identity and thematic interests.[5]

2 European Fission: American Fusion?

If one looks up HPS in the "European Reference Index for the Humanities" (ERIH), which sought to document the fields of research in 15 panels, one finds the following broad description[6]:

> History and Philosophy of Science covers the history of scientific disciplines, incl. medical and social science research and technology, as well as philosophical and social studies of science. In Europe, history and philosophy of science is not a well-defined discipline in institutional terms. Scholars publish their work in journals not always specifically identifiable as belonging to this field. However, philosophers of science are more likely to publish in general philosophy journals than are historians of science to publish in general history journals. Journals dealing with history and philosophy of the social sciences are included, as are journals dealing with the history and philosophy of logic, on the grounds that logic was a key area for philosophy of science.

Here it becomes clear that the English terminology can cause misunderstandings. While the English word "science" refers to the natural (and formal) sciences, the German term "Wissenschaft" includes the natural and social sciences and humanities, which ultimately also implies something programmatic. It is certainly not self-evident that the abbreviation HPS also includes cultural studies. Precisely the

[3]For example: "Philosophy of science without history of science is empty; history of science without philosophy of science is blind." (Lakatos 1971, p. 91)

[4]See the "Austrian classification of scientific disciplines" (Österreichische Systematik der Wissenschaftszweige), recommended by the FWF (Austrian Science Fund) for the disciplinary classification of project proposals: under the label "Geisteswissenschaften" (humanities), history and philosophy of science is seen as part of philosophy, while history of science as well as history of the social, the cultural, natural and the technical sciences is also classified under the label "historical sciences" ("Historische Wissenschaften").

[5]It is said that Kuhn was not happy as a member of the History Department in Berkeley, see Laudan (1990). Also, an Anglo-American specificity is the emergence of the Sociology of Science and Technology Studies (STS) and its (excluding or including) relation to HPS.

[6]ESF-ERIH Website: www.esf.org

European "methodological debate" that began at the end of the nineteenth century referred to the unity or diversity of science from an epistemological and philosophy of science perspective.[7] In any event, the case that Kuhn made in this connection is certainly striking. He did not argue for a unification of history of science and philosophy of science in one discipline to the benefit of both areas, but instead pleaded for a peaceful and productive coexistence in view of two different cultures of science – without the humanities being included.

> Few members of this audience will need to be told that, at least in the United States, the history and the philosophy of science are separate and distinct disciplines. Let me, from the very start, develop reasons for insisting that they be kept that way. Though a new sort of dialogue between these fields is badly needed, it must be inter- not intra-disciplinary. Those of you aware of my involvement with Princeton University's Program in History and Philosophy of Science may find odd my insistence that there is no such field. At Princeton, however, the historians and the philosophers of science pursue different, though overlapping, courses of study, take different general examinations, and receive their degrees from different departments, either history or philosophy. What is particularly admirable in that design is that it provides an institutional basis for a dialogue between fields without subverting the disciplinary basis of either.[8]

Against the backdrop of the success of Kuhn's *Structure of Scientific Revolutions*, this adherence to the duality of two areas of research (with the topoi of narrative/description vs. law-like general statements and different practice of reading) married to a call for constructive cooperation represents a modest variant of HPS voiced as the self-understanding of an "active historian of science".[9] Of course, Kuhn had to ask what was the added value of two such different fields joining in dialogue. The historians of science need philosophy as an indispensable tool since up until the end of the seventeenth century a large part of science simply was philosophy. According to Kuhn, the study of ideas was a genuine concern of philosophers, e.g., as for A.O. Lovejoy and Alexandre Koyré. Yet in the final analysis there remained doubts about the use and benefit of philosophy of science for the history of science. Even if there were important impulses (generally by neo-Kantian contributions or through Emile Meyerson or Léon Brunschvig), contemporary philosophy of science (e.g., C.G. Hempel) was seen as hardly relevant for the historian. Thus followed his call to historicize the theory of science – where the history of science is understood as concerned with the development of scientific ideas, methods and techniques and theory of science as a "general philosophy of science" (structure of scientific theories and theoretical concepts). In that way history of science – as well as the sociology of science – was to contribute to bridging the rift between philosophers of science and (natural) scientists by providing the necessary data and problems. By clearly privileging the historians over philosophers, Kuhn suggested "that the history and philosophy of science continue

[7] Stadler (2004).

[8] Kuhn (1976, p. 4). This paper is based on a talk Kuhn gave at the Michigan State University in 1968.

[9] Kuhn (1962/1970).

as separate disciplines. What is needed is less likely be produced by marriage than by active discourse."[10] (Here it becomes clear that the dual nature of HPS is expanded by the addition of sociology of science which together with the history of science could enhance the status of philosophy of science.)[11]

Different reasons for the separate development of history of science and philosophy of science in different directions since the 1930s were noted in 1990 by Larry Laudan when he noted, in opposition to Kuhn, that his interpretation of theoretical change systematically refused to take into account philosophical contexts.[12] "One can point to no classic essay or monograph which shows definitely why the call of Kuhn, Crombie, Conant, and Koyré (echoing the earlier call of Whewell, Mach, and Duhem) for historians of science to engage issues in the theory of knowledge has fallen on such deaf ears." Here a new argument emerges which brings into play both classical epistemology and historical epistemology and makes evident the necessity to explore how the conditions of the possibility of knowledge can be considered as part of a historical account of the dynamic of theories in concrete historical phases. According to Laudan, Kuhn was – against his own intentions – co-opted by the history of science camp. Here it is not possible to analyze in detail this indirect controversy but it is important to underline that the triangulation of this field – history, philosophy and sociology of (natural) science – spurred on the conceptual debates but certainly did not help to conclude them. It remains an open question how the social and historical aspect of science between can be systematized with epistemological consequences.

It seems that in the interwar years the European seeds of a historically and sociologically driven philosophy of science (dealing with nature, culture and society) were sown, but as a result of the largely forced emigration of scientists and scholars the transatlantic transfer and the intellectual shift towards analytic philosophy of science – with HPS becoming established in the English-speaking world – privileged the natural and formal sciences (logic, mathematics). There emerged a self-contained history of science that ultimately, as HPS, also led to the noted parallel emerging in philosophy of science and history of science. The contacts between North America and Europe during the Cold War resulted in a more limited research agenda. Only in the last two decades has the agenda of philosophy of science been expanded again through a growing interest in historical and sociological issues.[13] The European pluralism of development perspectives prior to World War II was followed by the American fusion in the model of HPS

[10] Kuhn (1976, p. 20).

[11] Compare with Hoyningen-Huene (1991, p. 43): "First, historiography of science provides the basis for both philosophy and sociology of science in the sense that the fundamental questions of both disciplines depend on the principles of the form of historiography employed. Second, the fusion of sociology and philosophy of science, as advocated by Kuhn, [. . .] consists essentially in a replacement of methodological rules by cognitive values that influence the decisions of scientific communities. As a consequence, the question of the rationality of theory choice arises, both with respect to the actual decisions and to the possible justification of cognitive values and their change."

[12] Laudan (1990, p. 50). Compare also with Laudan (1978).

[13] See Stadler (2010a) and Reisch (2005).

and HOPOS, which is primarily reflected in the respective programs and societies. This certainly did not imply a theoretical and methodological convergence in the sense of sharing epistemological interests and results. In other words, the North American fusion represents a combination of institutions which, in theory and practice – in spite of the numerous HPS programs and departments –, did not lead to cooperation and theoretical interaction. At the same time, the potential of the European integrative perspective from the interwar years was suppressed by the forced emigration and transformation of thematic interest – a potential which also included the origins of historical epistemology in Ludwik Fleck and the strong French tradition.[14] The American fusion remained prominent, while the European split between into continental and analytic theoreticians since 1945 was a result of political and scientific ruptures – resulting in a narrowed view of the relation of philosophy of science and the history of science.[15]

In the Boston Center for History and Philosophy of Science, founded by Robert C. Cohen and Marx Wartofsky in 1961, the sociological and historical tradition of European origin (Germany, England and France) was taken up and continued – pointing us retrospectively in the direction of historical epistemology and "science in context".[16] Looking back, one concrete example of a successful fusion of the history and philosophy of science remains the fragmentary life work of Edgar Zilsel (1891–1944) who due to his emigration and his early suicide in US exile was not able to bring to a conclusion his studies on the emergence of modern science (or on the notion of genius) that can be related to the contemporary discussions between Robert Merton and Joseph Needham.[17] Likewise first steps towards an experimental philosophy (of science) were taken by Arne Naess before World War II, which have only been taken up again and further developed in the last decade.[18]

3 On the Birth of HPS: Ernst Mach and His Legacy

If we would like to describe the Central European origins of HPS, then we could make no focus better focus than the physicist and natural scientist, philosopher and historian of science Ernst Mach (1836–1916).[19] He never ceased tostress the

[14]Within the limits of this paper, we cannot deal with the strong renaissance of the research on Fleck. Also, the forgotten dialogue of European philosophers with the French philosophy of science before World War II would need a more extensive research. See Nemeth and Roudet (2005) and Fleck (2011).

[15]Stadler (2010b).

[16]One of the last volumes of the series (vol. 263) of the "Boston Studies in the Philosophy of Science" with the programmatic title *Integrating History and Philosophy of Science* goes already in that direction (see also footnote 63).

[17]Zilsel (2000).

[18]Manninen and Stadler (2010).

[19]For an overview on Mach: Haller and Stadler (1988) and Wolters (1987).

historical dimension of all (natural) science and made a case for *Clio* as the main mentor of all scientific study. Already in his foundational early study of 1872 we read[20]:

> Classical education is essentially historical education. If this is true, then we have a much to narrow notion of classical education. Not the Greeks alone, the entire cultural life of the past, is important to us. Yes, there is a special cultural education for the natural scholar, which consists in the knowledge of the development history of his science. Let us not let go of history's guiding hand. History has made everything, history can change everything. We can expect everything of history . . .

This programmatic intention was already reflected in the titles of his main books, beginning with *The Science of Mechanics: a critical and historical account of its development* (1883). *Principles of the Theory of Heat*, to *Knowledge and Error. Sketches on the Psychology of Enquiry* (also see the alternative view later offered in Popper's *Logic of Scientific Discovery* (1934), and also in his late *Culture and Mechanics* (1915).[21] The lectures that he gave between 1895 and 1901 also document this historical orientation.[22]

- Development of the mechanical sciences. On psychology and logic of research
- History of acoustics and optics, viewed epistemologically
- History of the theory of heat and energy. Critical discussion on the teaching of physics
- History of the theory of electricity, viewed epistemologically. On several general issues of science
- Main epochs of science, viewed from an epistemological-psychological perspective. On several special issues of psychology
- The development of mechanics, viewed epistemologically. Psychology and logic of research
- The development of mechanics, discussed from a epistemological-critical perspective

In 1895, Mach received an offer which was to take him from Prague to the University of Vienna as a "Professor for philosophy, in particular history and philosophy of inductive science", the third chair of philosophy at the Institute of Philosophy.[23] This professorship had been initiated by the philologist Theodor Gomperz and his son, philosopher Heinrich Gomperz who – possibly on the

[20]Mach (1909, pp. 3f.).

[21]The essential writings of Mach are currently edited with introductions by different Mach scholars in the "Ernst Mach Studienausgabe" by the Berlin publisher xenomoi. (Editors: Friedrich Stadler, together with Michael Heidelberger, Dieter Hoffmann, Elisabeth Nemeth, Wolfgang Reiter, Jürgen Renn, Gereon Wolters).

[22]A selection of Mach's lectures, cited in: Mach (2011a, pp. XIII f.).

[23]On Mach's appointment see Mayerhöfer (1967).

basis of the model of Helmholtz[24] – were able to interest a philosophically and psychologically oriented natural scientist and mathematician for the so-called inductive sciences which at the time comprised the empirical natural sciences (physics, biology, physiology). Even though a serious stroke suffered by Mach in 1897 strongly impacted this last stop in his career – he had to retire in 1901 – one can speak of an exemplary institutionalization of the a fusion of history and philosophy of the natural sciences.[25]

In spite of his empiricist orientation Mach was wary of an exclusively inductive orientation, since he linked a fallibilist epistemology with a monist methodology by means of which he sought to overcome the dualisms of explanation and understanding, moving between poetic fantasy and the principle of economy (parsimony). In his late work *Knowledge and Error* he writes:

> From all that it emerges that the mental operation by which new insights are gained, which is usually called by the unsuitable name 'induction', is not a simple process but a rather complex one. Above all it is not a logical process, although logical processes may figure as auxiliary intermediate links. Abstraction and activity of phantasy does the main work in the finding of new knowledge. ... Since there is no adequate method to guide us towards scientific discovery, successful discoveries appear in the light of artistic achievement as was well known to Johannes Müller, Liebig and others.[26]

Mach concludes that the term "inductive sciences" is not justified for natural sciences. Mach can be seen as one of the pioneers of a HPS and of the historical turn in the philosophy of science – as has been attested to by the late Paul Feyerabend. By his equal treatment of both the contexts of discovery and of justification Mach laid the foundation for an integrated history and philosophy of science, long before Hans Reichenbach formulated the distinction of the contexts in 1938.[27] Moreover, he was aware of the important function of the then still underdeveloped conception of research heuristics (which since the 1970s, combined with greater focus on the context of discovery, has resulted in promising research)[28] and of the role of thought experiment as a major part of every process of inquiry and research.[29] One can find many common ideas in Mach and Ludwig Boltzmann (1844–1906) – including evolutionism, anti-metaphysics, empiricism: in consequence, the myth of the battle between the two prominent scholars and court councilors ("Hofräte") over atomism does not hold water. A thematic continuity becomes evident in his inaugural lecture on natural philosophy (1903):

---

[24] See Helmholtz (1921/1998). Helmholtz (1821–1894) was professor for physiology in Königsberg, professor for anatomy and physiology in Bonn, professor for physiology in Heidelberg and finally professor for physics in Berlin, but contrary to Mach, he was not on a chair for philosophy.

[25] On the reception of Mach: Stadler (1982).

[26] Mach (2011a, English trans. 1976, pp. 235–236).

[27] Reichenbach (1938). For the context from a contemporary point of view, see Stadler (2011).

[28] Simon (1977). For a recent overview: Schickore and Steinle (2006). See also Wolters (1986).

[29] Mach (2011b).

> Mach has so imaginatively elaborated that no theory is absolutely true, but also that there is hardly one that can be absolutely wrong, that instead each theory has to gradually be perfected just like organisms according to Darwin's theory. Given the fact that they meet strong resistance, what is not expeditious gradually falls away, while the expeditious remains and thus I believe that I am doing Prof. Mach the greatest honor when I contribute in this way what is mine to the further development of his ideas as far as it is in my power to do so.[30]

Here the skepticism shown by Boltzmann towards metaphysical and idealistic philosophy becomes manifest, a scepticism which was also reflected in his *Popular Writings* (1905) with contributions "On the Meaning of Theories" or "The development of methods of theoretical physics"[31] or even his entry "Model" in *Encyclopedia Britannica*.[32]

Before Moritz Schlick, the founder of the Vienna Circle, had taken over the chair for natural philosophy in Vienna, the philosopher Adolf Stöhr (1855–1921) had succeeded Mach as professor for philosophy of inductive sciences from 1911 to 1921.[33] Even though he himself covered the traditional philosophical canon, Stöhr created important impulses for a science-oriented philosophy with sociological and psychological aspects by establishing the Department of Philosophy and an Institute for Experimental Psychology at the adult education center (*Volkshochschule Ottakring*, *"Volksheim"*). Among his publications one finds titles such as an *Outline of a Theory of Names* (1889), *Guide to Logic in a Psychological Presentation* (1905) or *Psychology – Facts, Problems, Hypotheses* (1917) and he lectured on the logician Bernouilli, Hume's inductive logic and the probability calculation, psychological introduction to the study of sociology, Darwinism, Fechner as philosopher and psychophysicist, the empirico-criticism of Avenarius, experimental psychology and natural philosophy.

Yet it was only with the physicist-philosopher Moritz Schlick (1882–1936) that "scientific philosophy" was systematically introduced and elaborated in the context of the second natural science revolution (quantum physics and theory of relativity) and the linguistic philosophy of Wittgenstein in the Vienna Circle. There, however, the historical concern appears weaker in comparison to the philosophical one.[34] As opposed to the radical ambitions of a physicalist unified science and an *International Encyclopedia of Unified Science* associated with Rudolf Carnap, Otto Neurath and Charles Morris, Schlick continued to advocate – up until his untimely death – a "consistent empiricism" that supports not only a dualism of philosophy (clarifying the meaning of statements) and science (identifying facts), but also a parity of philosophy of nature and culture with ethical and aesthetic writings – even when in

[30] Boltzmann (1903).

[31] Boltzmann (1905).

[32] Boltzmann (1902).

[33] On his life and work: Stöhr (1974).

[34] Stadler and Wendel (2006a, b) and Engler et al. (2008).

his acceptance lecture (1922) he referred to his precursors.[35] "Almost all philosophy is natural philosophy" in Mach's and Boltzmann's sense for him, but Schlick did not merge philosophy of science and history of science:

> Philosophy following the strict thinking of the exact scientific philosophy, which views with contempt the opposition to natural science and its method of careful exploration of experience, is doomed to fail; no easier path to solving the greatest epistemological issues than through the individual disciplines. No royal path.

And he then continues:

> Orientation after exact thinking does not mean: limitation to the areas of philosophy which are obviously, directly related to natural science, physics and biology, but to treat all philosophical areas with the same love, but sub specie naturae, philosophy of history as well as epistemology, aesthetics as well as ethics.

As an interpreter of relativity theory appreciated by Einstein, Schlick revealed himself to be pluralist, even though from 1924 on, under Wittgenstein's influence, he was enthusiastic about the *Tractatus* philosophy. Both his book publications – e.g., *Lebensweisheit. Versuch einer Glückseligkeitslehre* (1918), *General Theory of Knowledge* (1918/1925), *Space and Time in Contemporary Physics. An Introduction to a general understanding of the theory of relativity* (1919), *Questions of Ethics* (1930) – and his lectures – which include: an introduction to natural philosophy, logic and epistemology, introduction to ethics, system of philosophy, questions of *Weltanschauung*, theory of relativity, problems of the philosophy of history, the problems of philosophy in context, historical introduction to philosophy, philosophy of culture and history – span a philosophy of history, culture and nature with a socio-critical intention covering all disciplines including epistemology. Even if no explicit philosophy of science as understood as HPS was aspired to, Schlick's project facilitated the contemporaneous emergence of a "logic of science" (Carnap) and philosophy of science (PSA). It is a strange but ultimately not surprising episode that the late Paul Feyerabend, after his massive criticism of all abstract-formal philosophy of science in *Against Method* (1970/1975), rediscovered Mach as one of the pioneers of the "historical tradition" in philosophy of science. The "philosopher from Vienna" praised the pragmatic-historical orientation with the following rhetoric:

> While Mach's criticism was part of a reform of science that combined criticism with new results, the criticism of the positivists and of their anxious foes, the critical rationalists, proceeded from some frozen ingredients of the Machian philosophy (or modifications thereof) that could no longer be reached by the process of research. Mach's criticism was dialectical and fruitful, the criticism of the philosophers was dogmatic and without fruit.[36]

[35] Moritz Schlick, "Vorrede" to: "Naturphilosophie". Manuscript in the Schlick Papers Nr.8 (Wiener Kreis Archiv Haarlem, North Holland).

[36] Feyerabend (1978, p. 202).

Feyerabend came up with seven (!) suggestions for an adequate practice in the philosophy of science[37]: (1) Against established views in the history of science. (2) Back to texts and sources. (3) Against simplification of established accounts. (4) Against pseudo-disputes and pseudo-problems in major philosophical issues such as positivism or realism. (5) Against purely philosophical problem solutions as mock attacks. (6) Against all too simple philosophies as opposed to complex historical processes. (7) Caution with philosophical fairytales. Seriously taking up these suggestions would surely amount to a radical renewal of the historical study of science and its philosophy – as had already been envisioned by other authors, even if their suggestions appeared too simple and incomplete.

4 History of Philosophy of Science (HOPOS) as a Good Example for HPS?

Already in the interwar period the question as to "and/or" of philosophy of science and history of science was being addressed in approximations. This can be shown consistently in Rudolf Carnap (1881–1970) who influenced significantly the development from "logic of science" to philosophy of science in his *Logical Syntax* (1934) with his triadic model of syntax, semantics and pragmatics. In the preface to his second edition he characterized the "logic of science" in the following way.[38]

> According to the current view this theory comprises, in addition to logical syntax, two further areas, namely semantics and pragmatics. While syntax is purely formal, i.e., only views the structure of linguistic expressions, semantics studies the relation of meaning between expressions and objects or concepts... Pragmatics also examines the psychological and sociological relationships between persons who use the language and the expressions.

Ten years later he defined "Science of Science" in the *Dictionary of Philosophy* (1944) as "the analysis and description of science from various points of view, including logic, methodology, sociology, and history of science." He also added references to "scientific empiricism" and the "Unity of Science movement" as a possible model for HPS.

It is a fact that the implementation and elaboration of the program remained incomplete for historical and theoretical reasons, but already since the turn of the century (1900) we can identify an abstract-theoretical tradition as opposed to a pragmatic-historical one, which in turn can be correlated with a philosophical "absolutism" vs. methodological "relativism".[39] In 1934 the Philosophy of Science Association (PSA) was established as an American-European initiative, with the journal *Philosophy of Science* was its publishing organ. This convergence

[37]Feyerabend, "Machs Theorie der Forschung und ihre Beziehung zu Einstein", in: Haller, Stadler, Ernst Mach (see footnote 20), pp. 461f.

[38]Carnap (1968, p. VII).

[39]Stadler (1997/2001).

between Central European logical empiricism and US-American (neo)pragmatism[40] addressed the tenuous relationship between philosophy and the (natural) sciences in the following mission statement.[41]

> Philosophy of Science is the organized expression of a growing intent among philosophers and scientists to clarify, perhaps unify, the programs, methods and results of the disciplines of philosophy and of science. The examination of fundamental concepts and presuppositions in the light of the positive results of science, systematic doubt of the positive results, and a thorough-going analysis and critique of logic and of language, are typical projects for this joint effort. It is not necessary to be committed to a belief that science and philosophy are or should be one, or else that 'never the twain shall meet'. If anything, there is to be expected a whole-some regard for the value of established science in furnishing a foil for philosophy and a check on its old extravagancies. This does not mean that even the best-established science may not be subject to most devastating criticism by analysis of its foundations. In fact, despairing of the philosophy of the schools, science has done it largely for itself. The theories of gravitation, atomicity, electro-magnetism, evolution, relativity and quanta have all arisen through drastic revisions of complacent fundamental 'truths'.

This wording strongly resembles Kuhn's dualism of history of science and philosophy of science and shows a similar division of tasks, placing hope in a productive dialogue while privileging the natural sciences (which is not so surprising). And it is also not surprising that the analytic philosophy of science that emerged in the cold war had lost the historical, political and value-related dimension. This development represented a historically conditioned (self-) censorship of the scientific community which was to be even more intensified by the criticism of insiders.[42] Here one could name as representative the much-cited texts by W.V.O. Quine, e.g., "Two Dogmas of Empiricism" (1951); another was Kuhn's *The Structure of Scientific Revolutions* (1962/1970), which was originally welcomed as a contribution to the *International Encyclopedia of Unified Science* by the editors Carnap and Morris; and finally P. Feyerabend's article "Against Method" (1970) which appeared in the *Minnesota* Studies *in the Philosophy of Science* (ed. by Herbert Feigl). In this connection one should also not forget to mention other manifestations of HPS: the five "International Congresses for the Unity of Science" (1935–1970) with Carnap, Morris, Neurath, the "Institute for the Unity of Science" (1947–1958) with Philipp Frank and Gerald Holton, the "Minnesota Center for the Philosophy of Science" (since 1953), the "Boston Colloquium and Boston Center for the History and Philosophy of Science" (since 1961 with the book series *Boston Studies in the Philosophy of Science*) with Robert S. Cohen and Marx Wartofsky (a pioneer of historical epistemology) as well as the "Pittsburgh Center for Philosophy of

[40]The members of the Editorial Board and the Advisory Board reflect this transatlantic cooperation and convergence between logical empiricism and neo-pragmatism. One gets a similar picture from the Advisory Committee of the International Encyclopedia of Unified Science (Carnap et al. (1938)).

[41]Malisoff (1934, p. 1).

[42]On the "cultural war on relativism" see Reisch (2005), especially the "Conference on Science, Philosophy and Religion" 1940–1968 as an expression of a xenophobic mentality since the entrance of the United States into World War II.

Science" with Adolf Grünbaum. Their scientific agenda became theoretically more oriented towards a non-historical "analytic philosophy of science". A dualism of philosophy of science and history of science emerged analogous to the distinction between context of justification and context of discovery.[43] Here, too, the historical reconstruction shows the loss of a historical potential for a comprehensive philosophy of science.

Against the backdrop of this development it was thus not surprising that the philosophy of science "returned" to the German-speaking countries of Europe after 1945 in the guise of a normative analytic philosophy of science which mainly promoted formal logic in the context of justification. The theory of science that became known as the known as "received view" or "standard view" of scientific theories was most pronounced in the German school of Wolfgang Stegmüller from 1959 on (theory structuralism) was, however, directly confronted in the 1960s with the renaissance of the historico-sociological tradition.[44] After a long period of dormancy there was a brief attempt to bring the latter back to life by the late Viktor Kraft (the Kraft circle included Bela Juhos, Paul Feyerabend as well as the Fulbright professor Arthur Pap in the academic year 1952/1953)[45] but it was very difficult for any modern philosophy of science to take hold in Austrian academe.[46] It was only with the establishment of new internationally oriented institutions outside of the universities by emigrants (Austrian College/Forum Alpbach, Institut für Wissenschaft und Kunst in Vienna, Institute for Higher Studies in Vienna) that are linked with names such as Carnap, Feigl, Frank, Menger, Popper, Lazarsfeld, and Morgenstern, that the state of the art of international research slowly began to influence philosophy at the universities – and then initially only in the provinces outside of Vienna.[47] Yet the early intellectual socialization of Feyerabend was decisive for his academic career which ultimately brought him back to the Vienna tradition with Mach as a pivotal figure of reference.[48]

Something like his plural interests happen to be reflected nowadays in two journals: on the one hand, the *Zeitschrift für allgemeine Wissenschaftstheorie – Journal for General Philosophy of Science* (from 1970), published in German and English, which also took into account the humanities, and the re-launched journal *Erkenntnis* (from 1975 on) in the tradition of analytic and formal philosophy of science following Carnap, Hempel and Stegmüller. Early attempts to reconceptualize HPS culminated in the "Colloquium on Philosophy of Science" in 1965 (preceded by the little noted 1961 History of Science Conference in London, which was also attended by Michael Polanyi, Popper and Kuhn) where the confrontation between Kuhn, Popper, Lakatos, Feyerabend, Toulmin and Watkins took place (and was continued in

[43] For an overview, see Stadler (2010b, 2012b).

[44] On this phase of slow intellectual reconstruction, see Stadler (2005, 2010a).

[45] On the life and work of Pap: Keupink and Shieh (2006).

[46] See Stadler (2012c).

[47] See Schorner (2010).

[48] Stadler and Fischer (2001).

the Proceedings).[49] A late consequence of this pluralist tendency was the founding of the International Society for the History of Philosophy (HOPOS) in 1996. Every 2 years this society organizes conferences in America and Europe and nowputs out its own journal which publishes articles on historical dimension of the philosophy of science since antiquity.[50] Its mission statement also promises "to include topics in the history of related disciplines and in all historical periods, studies through diverse methodologies" – which explicitly addresses the history of the sciences but leaves unaddressed how it relates specifically to the philosophy of science. The related *HOPOS* journal has formulated its goals in more concrete terms as:

> to provide and outlet for interdisciplinary work, increase the already unusually high level of participation from Europe and elsewhere in the history of the philosophy of science, raise the level of work in the history of philosophy of science by publishing scholarship that helps to explain the links among philosophy, science, and mathematics, along with the social, economic, and political context, which is indispensable for a genuine understanding of the history of philosophy.[51]

As shown by further initiatives that sought to take the connection and fusion of history of science and philosophy of science literally, which is evident in the group associated with an "Integrated History and Philosophy of Science" (abbreviated: &HPS), the scientific community did not just see historicizing and contextualizing philosophy of science as a sufficient integration of history of science.

5 Institutional Manifestations: Europe and North America in Comparison

If we compare the academic development with institutional manifestations, we get a different picture, namely, of a fusion of history of science and philosophy of science promoted by scientific societies under the banner of ideal-type programs and constructions. The *International Council for Science (ICSU)* founded in 1931, which presently has about 30 international and 100 national scientific societies under the auspices of the UNESCO formed the platform for the *International Union for History and Philosophy of Science* (*IUHPS*) which was founded in 1956, bringing together two societies that existed at the time: on the one hand, the *Division of Logic, Methodology and Philosophy of Science* (*DLMPS*, since 1955),[52] and on the other hand, the *Division of History of Science and Technology* (*DHST*, since 1949). Since 1971 there is a Joint Commission that is supposed to oversee joint activities and further forms of cooperation in keeping with the general goal "to

[49] On the LSE-conference 1965, see the Proceedings: Lakatos and Musgrave (1970).

[50] On the history of HOPOS: www.hopos.org

[51] *HOPOS – The Journal of the International Society for the History of Philosophy of Science* 1(2011), Statement of Policy, Cover page III.

[52] DLMPS und DHST websites: www.dlmps.org

establish and promote international contacts among historians and philosophers of science who are interested in the history and foundational problems of their discipline . . . "[53]

Yet already in the founding phase of the DLMPS there were differences over the subject matter that have not yet been resolved (between Alfred Tarski, Ferdinand Gonseth, Evert W. Beth, Stanislas Dockx, Patrick Suppes). Here the issue was the differentiation of logic and philosophy of science from pure mathematics and from philosophy. Moreover, there was originally some resistance from the history of science community to join the IUHPS.[54] Further diversification took place through the founding of several thematically oriented commissions. Thematically, a balance was to be achieved between "mathematical logic", "methodology of science" and "philosophy of science" which, however, varied strongly depending on place and local staff for the congresses held. "Technology" was also integrated in the DLMPS as in its sister division. Here the question arises to what extent the methodology of all sciences (of the deductive and the empirical sciences) could serve as a linking element – which might once again trigger a methodological debate that also takes account of the social and cultural sciences. The role of the formal sciences, in particular mathematics and mathematical logic in relation to philosophical logic would also be worth further discussion. To be sure, the program of the first congress at Stanford University featured five groups of empirical sciences: physics, biology and psychology, social studies, linguistics to the historical sciences. Yet the role of logic and mathematics as part of computer science and of cognitive science and biology offered a new forum for negotiating the changed conditions between the disciplines (which can only be hinted at here). A basic issue remains: the nature of the relationship and interaction of both divisions in connection with informal joint conferences and commissions.

HPS was embedded institutionally and professionalized also on a European level but this did not happen until relatively late: with the founding of the *European Philosophy of Science Association* (EPSA) in 2007[55] – modeled after the *Philosophy of Science Association* (PSA) and its journal – and a Research Network Programme for *Philosophy of Science in a European Perspective* (PSE).[56] Questions emerged about transatlantic cooperation on the one hand and cooperation between EPSA and the *European Society for the History of Science* (ESHS)[57] which had been organizing regular joint congresses along the North American model of cooperation between PSA and the History of Science Society (HSS), but with more unified themes and a broader understanding of science. (The latter was an institutional

[53] Mission statement in www.icsu.org

[54] Paulvon Ulsen, "The Birth of DLMPS. IUHPS/DLMPS", website: www.dlmps.org/history.html. Wilfrid Hodges, "DLMPS – Tarski's vision and ours". Opening address at the 14th Congress of Logic, Methodology and Philosophy of Science, Nancy, July 2011 (Hodges 2011).

[55] www.epsa.ac.at

[56] www.pse-esf.org

[57] www.eshs.org

initiative resulting more in synergies for both partner organizations than intellectual interactions.) Clearly it is easier to create an institutional foundation than to ensure an ongoing cooperation on the basis of a robust theoretical conceptualization of HPS (which given the unresolved issues and "camp mentality" of the players hardly comes as surprise). One must wonder whether the field's intellectual identity is not yet sufficiently developed or the contrasts between its constituents too difficult to bridge to allow for a joint program of science to be pursued. Is this a late confirmation of Kuhn's diagnosis?

6 Recent Research and Manifestations: "Integrated HPS"

Prompted by the deficits noted, members of some of the cited organizations created an international group aiming at an "Integrated History and Philosophy of Science" (&HPS).[58] The goal of this Euro-American initiative towards HPS as a discipline, at present still without a society and journal, is this:

> Good history and philosophy of science is not just history of science into which some philosophy of science may enter, or philosophy of science into which some history of science may enter. It is work that is both historical and philosophical at the same time. The founding insight of the modern discipline of HPS is that history and philosophy have a special affinity and one can effectively advance both simultaneously.[59]

Furthermore:

> What gives HPS its distinctive character is the conviction that the common goal of understanding of science can be pursued by dual, interdependent means. This duality may be localized in a single work. Or it may be distributed across many works and many scholars, with parts locally devoted just to historical or philosophical analysis. Intellectual history, for example, serves this purpose. What unifies this local scholarship into an HPS community is the broader expectation that all the work will ultimately contribute to the common goal. There is no distinct methodology that is HPS. Doing HPS does not confer a free pass to suspend the standards of one field to advance the other. It must be good history of science and philosophy, in that its claims are based on a solid grounding in appropriate sources and are located in the relevant context. And it must be good philosophy of science, in that it is cognizant of the literature in modern philosophy of science and its claims are, without compromise, articulated simply and clearly and supported by cogent argumentation.

Even though there is no uniform methodology, a consistent duality is postulated which is founded on the two-sided historical-philosophical character of the field. The conferences to date and their programs make clear that this agenda has only been realized in part. However, what is notable from the perspective of philosophy of science is the consistent attention to historical contexts: the heuristic principle represented by the abbreviation HS & PS as &HPS is being taken seriously.

---

[58] See the website of the latest and already forth conference in Athens: http://conferences.phs.uoa.gr/andhps/

[59] Folder of the &HPS Conference in Athens 2012 (see footnote 57).

It is not surprising that the description of "philosophy of science and the history of science" by one of the initiators of &HPS begins with an account of the attempt of philosophy of science to get history of science "back on board" after the breaks of the 1950s.[60] Successful historical episodes involving Whewell, Mach, Duhem, the circle of Emile Meyerson, and the ("left wing") of the Vienna Circle are cited before the distinction between the contexts of discovery and justification is critically addressed. This historical reconstruction provides the basis for the (re)unification of history and philosophy of science for a new generation of scholars as an ongoing symbiosis no longer stuck in the trenches of context distinction.

> The main point to emphasize is simply that the example of Friedman's *Dynamics of Reason* project makes vivid ..., the extent to which old assumptions about the alleged separation between history and philosophy of science are no longer valid. Common sense now holds that the marriage of history and philosophy of science is a stable and happy one.[61]

Another manifestation of this relatively new attempt at a "marriage" between history of science and theory of science can be found in the collection entitled *Integrating History and Philosophy of Science* (2012), which presents both general studies on the relations between the two fields in the USA since 1960 and case studies of exemplary contributions.[62] The editors suggest that it cannot be overlooked that the field was created mainly by philosophers nor that progressivist intellectual history (A. Koyré and Ch. Gillispie) served as its lowest common denominator until it was undermined by Kuhn with reference to Ludwik Fleck.[63] With Fleck, Wittgenstein and Kuhn, the contrast of descriptive work and normative assessment and the prominence of sociology of science as "sociology of scientific knowledge" (SSK) is said to have been consolidated. Scientific practice as opposed to theoretical development was focused on as the expression of contingent scientific cultures bringing into play the philosophical problem of relativism and its polarizing effects.[64]

What becomes clear here, given all the oppositions and interactions noted, is the fact that we are not dealing with a symmetric dialogue. The philosophical "community" is accentuating bridge building far more than historians had intended. This may be related to the former's normative and universalist orientation even though the strongly simplified "received view" was long discarded.[65] Analogous conclusions can be drawn about institutions. With few exceptions, initiative come more from philosophical than historical institutes. (This has also raised problematized the role of the history of science within the historical sciences.) Even if variants of a historicized or naturalized philosophy of science may be promoting

[60] Howard (2011, p. 66).

[61] Howard (2011, p. 67).

[62] Mauskopf and Schmaltz (2012).

[63] Mauskopf and Schmaltz (2012, introduction, pp. 1–10).

[64] See Bloor (2007).

[65] For an attempt to rehabilitate that view, see Lutz (2012). A similar exception are the books by Fritz Ringer, the most recent: Ringer (1997).

a rapprochement, the "marriage of convenience" once noted by Ronald Giere has not been strengthened. We find more temporary alliances and local knowledge than a unified field of research with the ideal of a discipline merging theory and practice. In the volume cited above we find at least a parity between philosophy and history of the sciences, albeit limited to the natural sciences. It would have made sense to follow Ernst Cassirer who explicitly addressed the philosophy of natural *and* cultural science.[66]

7 Excursus: "Decolonised HPS" and the Gender Perspective

It comes as no surprise that, with some delay, the legitimate question as to its Eurocentric perspective is currently being raised in HPS, now that the global and intercultural perspectives have become part of philosophy.[67] The dominance of Western rationality and the hegemony of Western history of science and philosophy of science as opposed to Arabic and Chinese culture prompted understandable objections to a privileged culture of science.

> The global consciousness needs new academic courses that transcend older intellectual tyrannies. Philosophy of science worldwide portray the development of science and its dynamic conceptual frameworks as a simple, linear, European development in which the rest of the globe played no role. This would be contrary to history. It is in fact false history. The proposal to decolonize education – on which most universities are agreed – requires, as a first step, the dismantling of false history, and the equally distorted philosophy of science which accompanies it.[68]

Cited as examples of such false histories were Kuhn's paradigm shift as based on the incorrect reception of Ptolemy and Copernicus, Newton's revolution as a misunderstanding of Indian mathematics, the fictitious Euclid as a pioneer of formal deductive proofs or the metaphysical conceptions of mathematics in Russell and Hilbert.

With a similar delay, the long neglect of the gender perspective is slowly being made up for.[69] The sociological-pragmatic turn in philosophy of science and the historicizing of its "history of dogmas" has not just raised the issue of the marginal or forgotten role of women in philosophy but also the relevance of an empirical-holistic philosophy whose contextual and relativist orientation encompasses the dimension of values (*inter alia*, social empiricism/epistemology in Helen Longino and Miriam Solomon). In this context Otto Neurath's anti-foundationalist and

---

[66] Ferrari (2012).

[67] This question had been addressed in the historiography of modern science and the scientific revolutions; see Collins (1998).

[68] Currently, HPS Curricula are developed which attempt to take the perspective of a "post-colonial HPS", see C.K. Raju, HOPOS-List, February 1, 2011.

[69] Potter (2006) and Grasswick (2011).

non-reductionist naturalism – a view that is associated with his well-known ship metaphor – is referred to: "... the relation of feminist thought to its discursive environment' can be grasped in terms of Neurath's boat, which cannot find a heaven safe from error but has to be repaired while out at sea."[70]

Janet Kourany argues in a similar vein for "A Philosophy of Science for the Twenty-First Century" when she draws a line from tradition of the Vienna Circle which was disrupted to the forced exile of many of its members, via the Cold War to present-day feminism.[71] Even though by now controversial studies dealing with the relation of naturalized philosophy of science, the "left Vienna Circle" and feminist philosophy of science have emerged, there still remain lacunae that need to be filled. Historicizing approaches and socio-cultural contextualizations are now being given greater emphasis, while the global perspective continues to be somewhat neglected. There is a dominance of "native English speakers" in the philosophy of science, which has rightly been seen as a problem for a multi-lingual and multi-ethnic European tradition.[72]

8 Summary: HPS Between Description and Construction

- Despite a variety of institutional manifestations HPS remains an open and dynamic research field in local contexts.
- HPS is inter- and trans-disciplinary, an orientation that is both exclusive (concerned with scientific knowledge) and inclusive (including the sciences *and* the humanities).
- In the context of HPS any relations of symmetry or asymmetry between philosophy of science and history of science are a contingent phenomenon.
- HPS can be conceived as the ideal type of a hybrid field of research on the basis of historical reconstruction.
- History of Philosophy of Science (HOPOS) is a special case that illustrates and manifests these finding.
- The theory and practice of HPS varies internationally and cannot be identified within a conventional disciplinary matrix: there is an ongoing process of cognitive identity formation.
- "Best practices" in teaching and research document productive interfaces between history of science and philosophy of science (esp. of an epistemological and methodological nature).
- HPS is ultimately embedded in concrete epistemic and scientific cultures that are informed by local conditions and academic politics within the spectrum of philosophical and historical institutes.

[70] Fricker and Hornsby in Fricker (2000, p. 3).

[71] *Philosophy of Science* 70 (2003), pp. 1ff.

[72] See Wolters (2014).

- It follows that HPS is to be regarded as a heuristic research program and interdisciplinary project with a future, since the disciplinary matrices of academic systems of knowledge are also subject to constant change. Whether a balance can ultimately be attained between the disciplines of philosophy and history is a question of local (self-)organization resulting from the policies pursued by the disciplines involved.
- The strong international presence of HPS programs in teaching and research allows us to conclude that the appeal of such an interdisciplinary, international research field as an ideal exceeds its concrete implementations.
- The HPS agenda will increasingly call into question the euro-centrist stance and will also have to give more attention to the gender perspective.
- Independently of this, the inclusive understanding of knowledge of the European tradition that is not limited to the last two centuries remains relevant for future developments.
- The tension between the empirical description and ideal-type construction of HPS reflects the basic openness of this dynamic field as well as the theoretical and conceptual problems that are purely discipline-related.
- The promise of HPS can only be borne out by concrete studies and joint projects – as building blocks of a future HPS (ideally with the input of sociology and technology) able to figure as an inter- and trans-disciplinary science.

References

Bloor, D. 2007. Epistemic grace. Antirelativism as theology in disguise. *Common Knowledge* 13: 2–3. doi:10.1215/0961754X-2007-007.

Boltzmann, L. 1902. Model. In *Encyclopaedia Britannica*, 10th ed. Reprinted in the 11th edition, vol. 10: 638–640.

Boltzmann, L. 1903. Ein Antrittsvortrag über Naturphilosophie. in: L. Boltzmann (1905), pp. 338–334. Cited after the edition by Engelbert Broda, Braunschweig: Vieweg 1979, pp. 199 f.

Boltzmann, L. 1905. *Populäre Schriften*, n. 18, Leipzig: Barth.

Carnap, R. 1968. *Logische Syntax der Sprache*. Vienna: Springer.

Carnap, R., C. Morris, and O. Neurath (eds.). 1938. *International encyclopedia of unified science*. Chicago: University of Chicago Press.

Collins, R. 1998. *The sociology of philosophies. A global theory of intellectual change*. Cambridge, MA: Harvard University Press.

Engler, F.O., M. Iven, and H.-J. Wendel (eds.). 2008. *Schlickiana*. Berlin: Parerga.

Ferrari, M. 2012. 'Wachstum oder Revolution'? Ernst Cassirer und die Wissenschaftsgeschichte. *Berichte zur Wissenschaftsgeschichte* 35(2): 113–130.

Feyerabend, P. 1978. *Science in a free society*. London: New Left Books.

Fleck, L. 2011. In *Denkstile und Tatsachen. Gesammelte Schriften und Zeugnisse*, ed. S. Werner and C. Zwettel. Frankfurt a.M: Suhrkamp.

Fricker, M. (ed.). 2000. *The Cambridge companion to feminism in philosophy*. Cambridge: Cambridge University Press.

Grasswick, H.E. (ed.). 2011. *Feminist epistemology and philosophy of science. Power in knowledge*. Dordrecht: Springer.

Haller, R., and F. Stadler (eds.). 1988. *Ernst Mach – Leben und Werk*. Vienna: HPT.

Helmholtz, W. 1921. *Schriften zur Erkenntnistheorie*. Comments by M. Schlick and P. Hertz. New edition by E. Bonk, Vienna/New York: Springer, 1998.
Hodges, W. 2011. DLMPS – Tarski's vision and ours. Opening address at the 14th Congress of Logic, *Methodology and philosophy of science*, Nancy, July 2011.
Howard, D. 2011. Philosophy of science and the history of science. In *Continuum companion in the philosophy of science*, ed. S. French and J. Saatsi, 55–71. London/New York: Continuum.
Hoyningen-Huene, P. 1991. Der Zusammenhang von Wissenschaftsphilosophie, Wissenschaftsgeschichte und Wissenschaftssoziologie in der Theorie Thomas Kuhns. *Journal for General Philosophy of Science* 22: 43–59.
Keupink, A., and S. Shieh (eds.). 2006. *The limits of logical empiricism. Selected papers of Arthur Pap*. Dordrecht: Springer.
Kuhn, T.S. 1962/1970. *Structure of scientific revolutions*. Chicago: University of Chicago Press. (Published in R. Carnap, C. Morris and O. Neurath, eds. *Unity of science. International encyclopedia of unified science*, vol. II, 2).
Kuhn, T.S. 1976. The relations between the history and the philosophy of science. In *The essential tension*, 3–20. London/Chicago: University of Chicago Press.
Lakatos, I. 1971. History of science and its rational reconstructions. In *PSA 1970, Boston studies in the philosophy of science VIII*, ed. R.C. Buck and R.S. Cohen, 91–136. Dordrecht: Reidel.
Lakatos, I., and A. Musgrave (eds.). 1970. *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press.
Laudan, L. 1978. *Progress and its problems. Towards a theory of scientific growth*. Berkeley: University of California Press.
Laudan, L. 1990. The history of science and the philosophy of science. In *Companion to the history of modern science*, ed. R.C. Olby, G.N. Cantor, J.R.R. Christie, and M.J.S. Hodge, 47–59. London/New York: Routledge.
Lutz, S. 2012. On a Strawman in the philosophy of science. *HOPOS. The Journal of the International Society for the History of Philosophy of Science* 2: 77–120.
Mach, E. 1909. *Die Geschichte und die Wurzel des Satzes von der Erhaltung der Arbeit*. Leipzig: Barth. original ed.: Prag: Calve 1872.
Mach, E. 2011a. In *Erkenntnis und Irrtum. Skizzen zur Psychologie der Forschung*, ed. E. Nemeth and F. Stadler. Berlin: Xenomoi. English trans.: Mach, E. 1976. *Knowledge and error – Sketches on the psychology of enquiry*. Trans. T.J. McCormack and P. Fouldes. Dordrecht/Boston: Reidel.
Mach, M. 2011b. Über Gedankenexperimente. In *Erkenntnis und Irrtum. Skizzen zur Psychologie der Forschung*, ed. E. Nemeth and F. Stadler, 193–210. Berlin: Xenomoi.
Malisoff, W.M. 1934. What is philosophy of science? *Philosophy of Science* 1(1): 1.
Manninen, J., and F. Stadler (eds.). 2010. *The Vienna circle in the Nordic countries. Networks and transformations of logical empiricism*. Dordrecht: Springer.
Mauskopf, S., and T. Schmaltz (eds.). 2012. *Integrating history and philosophy of science. Problems and prospects*. Dordrecht: Springer.
Mayerhöfer, J. 1967. Ernst Machs Berufung an die Wiener Universität. In *Symposium aus Anlaß des 50.Todestages von Ernst Mach*, Freiburg i. Br.: Ernst Mach Institut, 12–25.
Nemeth, E., and N. Roudet (eds.). 2005. *Paris – Wien. Enzyklopädien im Vergleich*. Vienna/New York: Springer.
Potter, E. 2006. *Feminism and philosophy of science. An introduction*. London/New York: Routledge.
Reichenbach, H. 1938. *Experience and prediction. An analysis of the foundation and structure of knowledge*. Chicago: University of Chicago Press.
Reisch, G. 2005. *How the cold war transformed philosophy of science. To the icy slopes of logic*. Cambridge: Cambridge University Press.
Ringer, F. 1997. *Max Weber's methodology. The unification of the cultural and social sciences*. Cambridge/London: Harvard University Press.
Schickore, J., and F. Steinle (eds.). 2006. *Revisiting discovery and justification. Historical and philosophical perspective on the context distinction*. Dordrecht: Springer.

Schorner, M. 2010. Comeback auf Umwegen. Die Rückkehr der Wissenschaftstheorie in Österreich. In *Vertreibung, Transformation und Rückkehr der Wissenschaftstheorie. Am Beispiel von Rudolf Carnap und Wolfgang Stegmüller*, ed. F. Stadler, 189–252. Vienna/Munich: LIT Verlag.

Simon, H.A. 1977. *Models of discovery and other topics in the methods of sciences*. Dordrecht: Reidel.

Stadler, F. 1982. *Vom Positivismus zur "Wissenschaftlichen Weltauffassung"*. Vienna/Munich: Löcker.

Stadler, F. 1997/2001. *Studien zum Wiener Kreis. Ursprung, Entwicklung und Wirkung des Logischen Empirismus im Kontext*. Frankfurt a.M: Suhrkamp. English edition: Vienna/New York: Springer 2001. Spanish edition: Mexico City/Santiago de Chile: Fondo de Cultura Economica 2010.

Stadler, F. 2004. Induction and deduction in the philosophy of science: A critical account since the methodenstreit. In *Induction and deduction in the science*, ed. F. Stadler, 1–16. Dordrecht: Kluwer.

Stadler, F. 2005. Philosophie – Zwischen 'Anschluss' und Ausschluss, Restauration und Innovation. In *Zukunft mit Altlasten. Die Universität Wien 1945 bis 1955*, ed. M. Grandner, G. Heiss, and O. Rathkolb, 121–136. Innsbruck: Studienverlag.

Stadler, F. (ed.). 2010a. *Vertreibung, Transformation und Rückkehr der Wissenschaftstheorie. Am Beispiel von Rudolf Carnap und Wolfgang Stegmüller*. Vienna/Munich: LIT Verlag.

Stadler, F. 2010b. History and philosophy of science. From Wissenschaftslogik (logic of science) to philosophy of science: Europe and America, 1930–1960. In *Vertreibung, Transformation und Rückkehr der Wissenschaftstheorie. Am Beispiel von Rudolf Carnap und Wolfgang Stegmüller*, ed. F. Stadler, 9–84. Vienna/Munich: LIT Verlag.

Stadler, F. 2011. The road to 'experience and prediction' from within: Hans Reichenbach's scientific correspondence from Berlin to Istanbul. *Synthese* 181: 137–155.

Stadler, F. 2012a. History and philosophy of science. Zwischen Deskription und Konstruktion. *Berichte zur Wissenschaftsgeschichte* 35: 217–238.

Stadler, F. 2012b. The Vienna Circle: Moritz Schlick, Otto Neurath and Rudolf Carnap. In *Philosophy of science: Key thinkers*, ed. J.R. Brown, 53–82. London/New York: Continuum.

Stadler, F. 2012c. Wissenschaftstheorie in Österreich seit den 1990er Jahren im Internationalen Vergleich – Eine Bestandsaufnahme. *Journal for General Philosophy of Science* 43(1): 137–185.

Stadler, F., and K. Fischer (eds.). 2001. *Paul Feyerabend – Ein Philosoph aus Wien*. Vienna/New York: Springer.

Stadler, F., and H.J. Wendel (eds.). 2006a. *Moritz Schlick Gesamtausgabe (MSGA)*. Vienna/New York: Springer.

Stadler, F., and H.J. Wendel (eds.). 2006b. *Schlick Studien*. Vienna/New York: Springer.

Stöhr, A. 1974. In *Philosophische Konstruktionen und Reflexionen*, ed. F. Austeda. Vienna: Deuticke.

Wolters, G. 1986. Topik der Forschung. Zur wissenschaftlichen Funktion der Heuristik bei Ernst Mach. In *Technische Rationalität und rationale Heuristik*, eds. C. Burrichter, R. Inhetveen and R. Kütter, 123–154. Paderborn-München-Wien-Zürich: Schöningh.

Wolters, G. 1987. *Mach I, Mach II, Einstein und die Relativitätstheorie. Eine Fälschung und ihre Folgen*. Berlin/New York: de Gruyter.

Wolters, G. 2014. Globalized parochialism, or: Is there a European philosophy of science? In *Philosophy of science in Europe – European philosophy of science and the Viennese Heritage*, ed. M.C. Galavotti, E. Nemeth, and F. Stadler. Vienna/New York: Springer.

Zilsel, E. 2000. *The social origins of modern science*. Foreword by J. Needham. Introduction by D. Raven and W. Krohn, ed. D. Raven and R.S. Cohen. Dordrecht: Kluwer.

Index

M.C. Galavotti et al. (eds.), *New Directions in the Philosophy of Science*, The Philosophy of Science in a European Perspective 5, DOI 10.1007/978-3-319-04382-1,