The European Philosophy of Science Association Proceedings
Series Editor: Friedrich Stadler

Vassilios Karakostas
Dennis Dieks *Editors*

EPSA11 Perspectives and Foundational Problems in Philosophy of Science

epsa
European Philosophy of Science Association

Springer

The European Philosophy of Science Association Proceedings

Volume 2

Series Editor

Friedrich Stadler, *Wien, Austria*

For further volumes:
http://www.springer.com/series/8888

The European Philosophy of Science Association Proceedings

These volumes collect a selection of papers presented at conferences of the European Philosophy of Science Association. The volumes provide an excellent overview of the state of the art in philosophy of science as practised nowadays in different European countries.

Vassilios Karakostas • Dennis Dieks
Editors

EPSA11 Perspectives and Foundational Problems in Philosophy of Science

Springer

Editors
Vassilios Karakostas
Department of Philosophy
and History of Science
Faculty of Sciences
University of Athens
Athens, Greece

Dennis Dieks
Inst. History & Foundations of Science
Utrecht University
Utrecht, The Netherlands

ISBN 978-3-319-01305-3 ISBN 978-3-319-01306-0 (eBook)
DOI 10.1007/978-3-319-01306-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013947738

Printed on acid-free paper

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

Contents

Part I Philosophy of Science: Formal Philosophy of Science

Part II Philosophy of Science: Idealization, Representation and Explanation

Part III Philosophy of Science: Realism, Anti-realism and Special Science Laws

Part IV Philosophy of the Physical Sciences: Philosophy of Quantum Mechanics

Part V Philosophy of the Physical Sciences: Perspectives on Spontaneous Symmetry Breaking

Part VI Philosophy of the Physical Sciences: Philosophy of Space and Time

Part VII Philosophy of the Physical Sciences: From Physics to Metaphysics

Part VIII Philosophy of the Physical Sciences: Philosophy of Chemistry

Part IX Philosophy of the Life Sciences

Part X Philosophy of the Life Sciences: Biological Knowledge and Structural Realism

Part XI Philosophy of the Cognitive Sciences

Part XII Philosophy of the Social Sciences

Introduction

This volume contains a selection of the papers presented at the third conference of the *European Philosophy of Science Association* (EPSA) held in Athens, Greece, 5–8 October 2011. EPSA was founded in 2007 with the aim of stimulating the study of the philosophy of science in Europe, in a worldwide context. An important instrument for achieving this goal has been the organization of biennial conferences and the subsequent publication of the best delivered papers. The third EPSA conference (EPSA11), and this volume, continue this ongoing tradition that has been successfully established with the previous two conferences, EPSA07 in Madrid and EPSA09 in Amsterdam.

The papers collected here offer a representative sample of the best work in contemporary philosophy of science as currently practised in Europe and elsewhere. Refereeing and selecting papers for presentation and publication is a difficult task when so many good papers are submitted for consideration. The selection process has been extensive and took place in two stages. Following a call for papers, the Programme Committee of EPSA11, chaired by Kristina Rolin and Dennis Dieks and consisting of 31 experienced members, worked hard to choose a high-quality and balanced set of papers to be presented at the conference (162 contributions were selected out of more than 400 submissions). After the conference, the proceedings editors went on to make a further selection among the papers that were delivered. The selection was based on the evaluation reports of the initially submitted extended abstracts, with originality, significance, clarity and diversity of topics as the most important criteria. A severe reviewing process followed. Those papers that were finally accepted for publication have in almost all cases been revised in light of the comments and suggestions supplied by the referees. There are thus good grounds for the claim that the 40 papers included in this volume provide an excellent sample of the current state of the art in philosophy of science.

The outstanding quality of the papers and their broad spectrum of topics reflect the mission and ambition of EPSA. Our young association is in a state of further development and, in accordance with this, the present volume introduces a novelty in relation to the past practice of EPSA proceedings: instead of assembling papers alphabetically, according to author's initials, the editors have organized the volume in thematic terms. In this way we have tried to meet the reader's natural wish for a structured book, even in the case of conference proceedings.

There remains the pleasant task of expressing thanks where these are due. First and foremost, we want to thank the Local Organizing Committee of EPSA11, chaired by Stathis Psillos, for all the work they have put in making the conference the great success it was. Further thanks are due to the Department of Philosophy and History of Science at the University of Athens, and especially its chairman Costa Dimitrakopoulo, for providing financial and administrative aid. Thanks also to Henk de Regt for his support during the early stages of the editorial process. Our sincere gratitude extends to members of the Programme Committee, and a substantive number of external referees, who helped us enormously in evaluating the papers submitted for publication in this volume. Finally, we want to take the opportunity to thank Ties Nijssen, Christi Lue and Madhuriba Subarayalou at Springer for their careful work in the final compilation of the volume. The two editors themselves see this volume as the tangible finished product of 2 years of effective and pleasant collaboration.

Athens Vassilios Karakostas
Utrecht Dennis Dieks
April 2013

Part I
Philosophy of Science: Formal Philosophy of Science

Evidence, Argument and Prediction

Nancy Cartwright

Abstract In this paper I propose a theory of evidence – which I call *the Argument Theory* – for domains where it is appropriate to demand high standards of rigor, explicitness and transparency, as in evidence for scientific conclusions and especially for evidence-based policy, which is where the need for such a theory first became apparent to me. I then apply the Argument Theory to answer a question that is too seldom asked, and never properly answered, in evidence-based policy where randomized controlled trials (RCTs) are taken as the 'gold standard' for evidence for predicting policy effectiveness: What does it take to makes positive RCT results evidence for policy predictions? The answer it turns out is quite a lot: information is required both about the causal role of the policy in the local circumstances and the helping factors required for it to work there.

1 The Context and the Problem

This paper is about evidence, specifically about evidence for *effectiveness predictions*: predictions that a well-described programme, policy or treatment will work for us, i.e. that the programme will result in an improvement in a well-specified outcome if we were to implement it in a targeted situation in a specific way – the way we would in fact implement it. Evidence-based policy advocates have invested a great deal of effort over the last few years in evaluating and providing warehouses for storing what they offer as evidence for hypotheses of this form in various areas of concern, warehouses to be visited by 'ordinary' policy makers and analysts. There includes for instance the Cochrane Collaboration for medical

N. Cartwright, LSE and UCSD (✉)
Department of Philosophy, Logic and Scientific Method, LSE, Houghton Street,
London WC2A2AE, UK
e-mail: n.l.cartwright@lse.ac.uk

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_1,

studies, the Campbell Collaboration for general social policy, the US Department of Education's What Works Clearinghouse, the George Mason University Centre for issues in criminology and the greater London Authority's new Project Oracle for 'Understanding and sharing what really works' against youth violence.

These warehouses advertise that they store programmes that 'work' to produce targeted results. We as philosophers know to be wary of sloppy language like that. What they store are programmes for which there is very good evidence that they work somewhere, and, if we are very lucky, in a few somewheres. The warehouse keepers police certain kinds of scientific studies, studies that aim to establish causal connections between a programme and a targeted outcome. Programmes that make it onto the shelves in the warehouse are ones that have been tested in what the warehouse regulations regard as very good studies. In particular the warehouse purchasing rules strongly favour RCT study designs – that is, randomized controlled trials.

What an RCT can evidence directly is that the programme worked there, then, in the study population. What makes that evidence for the effectiveness claim of concern to policy analysts: 'It will work here, now, as we would implement it?' What does it take for the RCT result to play a part in a support structure that argues for the truth of the effectiveness prediction? That's my question.

I propose the same answer I urge for claims in any domain where the demands for rigor and explicitness are high, as in warranting conclusions in science or for evidence-based policy, namely what I call *the Argument Theory of Evidence*: conclusions are warranted by good arguments, arguments that are both valid and sound. It is surely trivial to remark that a conclusion is warranted by a good argument. But this reminder helps underline two important facts that are not currently at centre stage in discussions about evidence for effectiveness predictions:

1. Evidence is a 3-place relation: e is evidence for h *relative to* a specific argument A for h. Failing the rest of the premises in A, or relative to a different argument A′, the very same fact, e, can be totally irrelevant to the very same hypothesis h.
2. Arguments are like chains: they are only as strong as their weakest premise. Focusing on the argument forces the premises to the fore. Often it is just the ones that aren't generally stated that turn out to be most dicey.

2 The Argument Theory of Evidence

2.1 The Theory and the Reasons for It

What is evidence? More specifically, under what conditions is one empirical claim *e* evidence for a second empirical claim *h*? I note from the start that evidence is not a natural kind. There is no 'correct' theory of what evidence is, as there might be a correct theory of what an electron is. When this is the case our account of what makes for a good theory should be responsive to what needs the theory addresses. The theory I propose started as a theory of 'evidence for use', in particular for use in making reliable predictions about what results will be produced by actions we consider taking. The Argument Theory is not confined to this context, however, but should fit anywhere we face the same needs.

A central problem I see everywhere that evidence-based policy is on the tapis is that the way the term ‘evidence’ is usually used lets in far too much. And it does so while at the same time purporting to be very restrictive by subscribing to the highest standards of rigour. In response my theory of evidence is demanding. That is because I agree with a common supposition. It is commonly – and I think reasonably – supposed that

Desideratum
A piece of evidence for a hypothesis should speak for the truth of the hypothesis.

It is with this in mind that I offer stringent criteria. I want criteria such that, once a fact meets those criteria, we should be happy to allow it to weigh in.

Here is what the *Argument Theory* demands of evidence:

> A well-established empirical claim *e* is evidence for hypothesis *h* relative to a good argument A (or A, A′, A//” . . .) if and only if e is a premise in A, which is itself a good argument for *h* (or, is a premise in A′ which is a good argument for a premise in a good argument A for *h*, etc.), where a good argument has true premises and is deductively valid.

The Argument Theory is akin to Clark Glymour’s bootstrapping theory of confirmation (Glymour and Stalker 1980) in which we bootstrap from evidence to hypothesis using background assumptions and inductive logic. On the Argument account, given the other premises (which are like Glymour’s background assumptions), h follows deductively from e. For Glymour, by contrast, the conclusion we derive from e and the background assumptions is an *instance* of h. Then we must use inductive principles to get from ‘instance of h’ to h.

For me h itself is the fixed point that we wish to arrive at. The Argument Theory requires that we do so by a good deductive argument. So I need far stronger background assumptions than Glymour. This, I urge, is all to the good. Science and evidence-based policy gain their high status in large part because they lay claim to being rigorous, public and explicit. These were the demands of Popper and of the Positivists and ones that we should insist on adhering to. There is in principle no objection to inferring h from instances of h in particular cases – so long as it is clear what it is about h and these instances that warrant this inference in this case. Are all the instances the same always? Are they at least all the same in this situation? Does the instance in question have special features that make it characteristic, so that if it holds, h holds? Or . . . ? The Argument Theory demands that the assumption that warrants the inductive leap be explicit in each case so it too can be subject to scrutiny. That’s because hiding what it takes for the conclusion genuinely to follow from the evidence is both morally and intellectually culpable in any enterprise that sails under the flag of science or of evidence-based policy.

Two parallel lines of defence support the Argument Theory, one ontological, the other epistemological. That’s because evidence is Janus-faced. On the one hand it has to do with truth and truth trackers: with what facts of Nature there are and what other facts can ensure they obtain. I should note that here I take a generous view of what the facts of Nature include. In particular, facts can be expressed by general claims, like Maxwell's equations or the claims of general equilibrium theory in economics, as well as by singular claims, like 'The cat is on the mat.' On the other hand, evidence has to do with our attempts to arrive at truths: with our hypotheses about what facts obtain and the further hypotheses that provide warrant for them. The two lines of reasoning are obverse sides of the same coin, one expressed – to use Carnap's terminology – in the material mode, the other in the formal mode.

Begin with the material mode. Some facts or sets of facts are sufficient for others: if the first obtains, the second cannot fail to obtain. One fact, f_e, is evidence for a second, f_h, then if f_e is a necessary member of a set of facts sufficient to ensure f_h obtains. Note that 'sufficient to ensure X obtains' is not the same as 'brings X about'. It means just what it says: the one set cannot obtain if the other fails.

If you were brought up in the tradition of Hempel and Nagel you may well be more comfortable with the formal mode version of the parallel lines of defence. Evidence for a claim is supposed to contribute to warrant for the truth of the claim. What contributes to warrant for the truth of a claim are reasons, and what makes some claim e a reason for another h is that e figures in a good argument for h. 'Good' here = valid and sound; the premises are true and the conclusion genuinely follows from them. Deduction provides a clear sense to what it means for a conclusion to follow from a set of premises. It is the formal mode counterpart to one set of facts being sufficient in Nature to ensure that a second obtains.

Beware the formal mode though. We are looking for a formal mode counterpart of the relationship in Nature where one set of facts is sufficient for another to obtain. Then evidence can satisfy the Desideratum that a piece of evidence for a hypothesis genuinely speaks for its truth. That is the sense of 'warrant' involved in the formal mode account of evidence. Alternatively 'warrant for h' sometimes means 'justifying a belief in h'. That is not the sense at stake here. Belief is an attitude or an action, and, I would argue, there is no context-independent sense of justification for it. Whether it is justified to hold a belief in h depends on what is consequent upon believing it. Will God send me to hell for it? Will I build a bridge supposing h is true, which bridge will fall down if h is false? Will I teach it to my graduate students who might then win a Nobel Prize by taking it as the basis for their research or alternatively, fail to get their PhD because their research went nowhere? Still what I think about justifying belief is an aside since belief is irrelevant to my topic.

Evidence in the sense supposed in the evidence-based policy literature and in the sense required for establishing scientific hypotheses has nothing to do with belief. It has to do with the truth of empirical claims and with what facts ensure that truth. So inductive logics and subjective probabilities have no place in the characterization of evidence for these purposes. Of course they may, if you believe in them, play a legitimate role when it comes to our estimates of whether one claim is evidence for another.

The demand that an evidence claim figure in a good argument – both valid and sound – may seem excessively strong. I actually make a stronger demand. Not only should there be a good argument from e to h if e is evidence for h, but we should not count e as evidence until that argument is displayed. I sometime express this in the slogan ‘It’s not evidence till there’s evidence it’s evidence.’ C.G. Hempel’s account of explanation also demanded validity and it also majored on deductive arguments. Hempel though allowed that many good explanations in science are enthymematic, in particular they are often not completely laid out. When it comes to evidence for scientific claims or policy predictions, I think it can be a bad mistake to allow this. Both science and evidence-based policy get their status in part from their claims to rigor. As a way to ensure rigor nothing beats laying out the arguments and looking to see how good they actually are.

2.2 Some Objections and Answers

There are a few objections philosophers may have right away to the Argument Theory of evidence. None, I urge, undermines the account.

- On this account of evidence we never know that a claim is evidence because that would require knowing that the claim is a necessary part of a good argument. To know the argument is good you need warrant for the other premises. To warrant those premises you need good arguments; to warrant that these arguments are good you need warrant for the premises in them. Etc, etc. That does not seem to me a problem: it’s what good honest evaluation requires. Of course we stop somewhere; we have to. In the best of cases we stop with claims that can be taken as well established. To the extent that our stopping points are not ones we can take for granted, to that extent we should be cautious about our supposition that a proffered evidence claim really is evidence after all. We know from Otto Neurath that in reasoning we are like sailors who must repair our boats at sea without ever putting in to dry dock to build from firm foundations. So we will always have to trust to some claims we take as true, at least for the nonce. But we should not make our situation worse by neglecting our arguments: without laying out all the premises in all the arguments we don’t know how leaky our boat is.
- The Argument Theory implies a number of what might be thought oddities, to all of which I have the same answer. Yes these facts are indeed evidence but they are not usually very useful pieces of evidence for us.
  - Anything true is evidence for a logical truth since anything – any claim at all – is a premise in a good argument for a logical truth. Yes, and so, I maintain, it should be. Anything is evidence for a logical truth. Still I wouldn’t advise spending much to buy information about other facts to warrant a logical truth. If you know a claim is a logical truth you don’t need to buy information about other facts to warrant the claim. And if you don’t know that the claim is a logical truth, you will have trouble warranting that the claim is implied by the fact you buy. Still, if I don’t know h is a logical truth but I am assured that if e then h, then e is surely worth learning.

  - *A&B* is evidence for A; *A* > *B* & *A* is evidence for B; etc. Yes, they are. But we know that conclusions of arguments are no more warranted by the argument than the premises, so we won't be led astray here in evaluating the warrant for the conclusion.
  - Everything is evidence for itself. That's ok. Any claim does speak for itself. Again though, we know that conclusions of arguments are no more warranted by the argument than the premises, so we won't be led astray here either.

- The Argument Theory employs a flawed theory of relevance. It lets in as evidentially relevant just the kinds of things philosophers have been at pains to rule out. Consider the canonical example: 'John Jones takes birth control pills.' Surely this is not evidence for his non-pregnancy. But I think, to the contrary, that it is excellent evidence:

  1. Nobody who takes birth control pills gets pregnant.
  2. John Jones takes birth control pills.

  Therefore: John Jones does not get pregnant.

  Given (1), (2) speaks – and speaks compellingly – for the truth of the claim that John Jones does not get pregnant. What better basis could the truth of this claim have? To suppose that John Jones's taking birth control pills is not evidence for his failure to get pregnant is to confuse the task of providing evidence that a fact obtains with the task of explaining why it obtains.
- If all arguments are deductive then on the Argument Theory

  - There can't be both evidence for a claim and evidence for its negation since evidence claims must be able to participate in good deductive arguments and there can't be good deductive arguments for a hypothesis and its negation. That's okay too. It can still be reasonable to say 'We have evidence for h and evidence for not-h' when there are results that can figure in plausible arguments for h and results that can figure in plausible arguments for its opposite. What matters is that we recognize that the results only count as evidence relative to some good argument so that we don't just let the result weigh in without commitment to the existence of these arguments.

    Of course if there are good arguments that are not deductive and hence the truth of the premises does not guarantee the truth of the conclusion, then it can be literally true that there is evidence for both h and evidence for not-h on the Argument Theory of evidence. But that is as it should be.

  - There can be no evidence for false claims. As soon as you know there is evidence for h by lights of the Argument Theory, you know that h is true. But that seems to me no problem. The problem is coming to know that e is evidence for h. This is a serious job and one of my concerns with the evidence-based policy literature, as I shall explain tomorrow, is that it does not take the job seriously enough, while all the while boasting that that is just what it does.

> Although I don't think our ordinary locutions count for much in efforts like mine here to make precise an everyday concept like evidence so that it can serve specific scientific purposes, I'll just note that often we do use the term 'evidence' in a way that supposes that there's no evidence for false claims. If I am accused of cooking the books or murdering Ackerly, I might very well respond, "But you couldn't have evidence for that. I didn't do it."

2.3 An Alternative Account of Objective Evidence and Why I Do Not Adopt It

My insistence that in science and policy we want a sense of evidence in which evidence for a hypothesis speaks for its truth echoes views of Sherrilyn Roush, who has done a great deal of very instructive thinking about evidence. In her book *Tracking Truth* (Roush 2005), Roush links a theory of evidence with her theory of knowledge – where the latter has to do with what we are entitled to claim for ourselves as knowledge. Her very first sentence in the chapter 'What is Evidence? . . . ' is on the knowledge side: 'It is a truism that the better one's evidence for a claim p the more likely one is to have knowledge that p.' [p. 149]. But like me Roush is keen to keep the enterprises of theory of knowledge and theory of evidence separate:

> . . . the notions of evidence that I am aiming for are objective in the following sense. That e is evidence for h is understood as holding in virtue of a factual relation between the statement's being true and the statement h's being true, not in virtue of anyone's believing that this relation exists. [p. 156]

Her basic idea is this: 'Intuitively, good evidence for a hypothesis is a discriminating indicator of the truth of the hypothesis,' [p. 154] where 'discriminating indicator' means some appropriate probabilistic analogue of 'h is true if e is true and false if e is false'.

Formally Roush's account of evidence requires that for good evidence:

- P(h/e) be high.

In order to satisfy what she calls the *leverage condition*, Roush in addition requires:

- The likelihood ratio [P(e/h)/P(e/−h)] be greater than 1.

Moreover it is highly desirable that

- P(e) be high.

There are a number of reasons that I do not adopt Roush's account, hinging primarily on the fact that it is still too much of a hybrid between a theory of evidence and an account of how to justify our claims to knowledge.

Where our accounts part company at the start is over what Rush calls 'Bayesianism'. For her this does not mean a subjective interpretation of probability. Rather –

'the Bayesian makes the idealizing assumption that all statements of the language in question possess probabilities. This is in contrast to the approach of classical statistics in which it is denied, for instance, that hypotheses have probabilities.' [p. 155] For the objective notion of evidence that Roush and I both have in view, though, it cannot be probabilities of statements that matter but rather probabilities of facts. I do not see that there generally are such probabilities. Probabilities for facts arise from chance set-ups, which are a special kind of nomological machine (Cartwright 1999), and while nomological machines are not all that rare, those that count as chance set-ups appear to be a small subset.

Then I disagree with each of her conditions in turn.

- P(e) is high. Roush insists on this in a debate about whether evidence should be surprising, which many, Bayesians especially, require. Her discussion at this point repeatedly refers to degrees of belief despite the fact that she means to be embarked on an objective theory of evidence. And I think that's a clue. If we are thinking about a license to 'accept' h, there are a variety of reasons to value observing consequences of h that were not expected beforehand: like worries about accommodation rather than novel prediction, or the demand that h have content that goes beyond summarizing what's already known.
- Notice I say here 'expected' – that has to do with subjective probabilities which are not relevant to the objective notion of evidence. On the objective side, I urge that e should be true, not objectively probable. High probability of e only comes in as a demand when we consider whether we should 'accept' that e is evidence.

  Consider an example where we might all be willing to suppose there are objective probabilities. We have three coins:

  – For C(1), P(h) = .2
  – For C(2), P(h) = 1 . . . .it is two-headed.
  – For C(3), P(h) = .2

  Imagine that the following chance-se-up is in place from time t(1) through time t(3):

  – At t(1), flip coin1
  – At t(2), if C(1) = h, at t(2) flip C(2)
  At t(2) if C(1) = t, flip C(3)
  – At t(3) either h occurs on C(1) or either heads or tails on C(2).

  Now consider e = 'C(1) = h at t(2)' and h = 'heads occurs at t(3)'. P(e) = .2. That is low. But e is compelling evidence for h. What I want to underline is that it is compelling evidence not despite its low probability but regardless of its probability. It would be evidence no matter what its probability. Even though e has an objective probability, that objective probability is irrelevant to its status as evidence. This claim is true in general I maintain.

What I would say about e is this: ‘C(1) = h at t(2)’, if true, is evidence not for h but for h’ = ‘At t(2) the objective probability of heads at t(3) is .2’. This I think is the right thing to say and it is what follows on the Argument Theory.

- The likelihood ratio is high. This is in aid of leverage. Roush tells us: ‘. . . evidence provides leverage on the truth of claims about the world. Specifically, knowing that the evidence statement is true is usually a lot easier than knowing that the hypothesis statement is true, and we use the former to help us make progress on the latter where we could not have made progress directly.’ [p. 158] Damien Fennell and I have elsewhere (Cartwright and Fennell 2009) explained problems we have with thinking the likelihood ratio can provide leverage in the way Roush wants. I won’t rehearse those worries here but rather make a more general point. I don’t see how to justify any condition that demands leverage in this sense for an objective notion of evidence. Leverage clearly makes sense when we are in the business of justifying our claims to knowledge or trying to estimate what to expect in the future. Suppose e, if true, is evidence for h. There is no point in spending a lot of money to learn whether e is true or not as an aid to deciding whether h is true when it is a lot cheaper just to learn h directly. But that has nothing to do with whether e is evidence for h or not.
- P(h/e) is high. Suppose this is so and P is an objective probability and e is true. Then the objective probability of h is P(h) = P(h/e) and on the argument account e is good evidence for this – and that is so whether P(h) is high or not. For Roush it is also evidence for h. One could make this stipulation as part of an objective account of evidence but I think it is misleading. We don’t have evidence that h will obtain, just that it can, or might or might well; more precisely, that it has probability P(h) of obtaining. There may be no harm in adding Roush’s requirement to the argument account but it will mean that there can be good evidence – in the fully objective sense – for false h’s, not just evidence we mistakenly thought was good. Evidence does not provide the same assurance as it does on the basic Argument Theory.

  Also, note that if it is added as an allowance on the Argument Theory it would play a different role than in Roush’s. For Roush this is what secures the relevance of e to h. On the Argument Theory, that is secured by arguments linking e and h. And that demand should be enforced here as well. We should still demand a good argument – valid and sound – for the claim that P(h) = φ.

- There is one other feature on which Roush and I differ but not, I think, disagree. That is on *discrimination*. For Roush e should track h; bracketing issues about probabilities, e should be true if h is. The Argument Theory requires only that h be true if e is. One could perfectly well add this. ‘Evidence’ even ‘objective evidence’ is not a natural kind with a fixed criteria or a fixed extension. I do not wish to opt for this stronger notion since it is far stronger than what seems supposed in the evidence-based policy literature and in the bulk of scientific cases I am familiar with. In particular it would undercut the claim that positive results in ideal RCTs are evidence for causal claims since positive results imply a causal connection between treatment and outcome but negative results do not show there is none.

- I also have a worry about probabilistic characterizations of evidence like Roush's even when the topic is not objective evidence but rather our entitlement to hold some cognitive attitude to a hypothesis or to use it in some way: probabilistic characterizations put the cart before the horse. Subjective probabilities, at least when we employ them in serious decision making, should have reasons behind them. Like what? Conditional probabilities generally play an important role, like P(h/e). How do we set that? One standard way is look to see if e is evidence for h and how strongly it speaks for h's truth, then set the probability of h given e accordingly. But to do that, we need some independent way of characterizing evidence that does not depend on our subjective probabilities.

3 What Makes RCTs Evidence for Effectiveness?

I have rehearsed the Argument Theory of Evidence because it can provide us with an answer to this question and an answer that matters to getting our predictions right in evidence-based policy.

The current evidence-based policy literature rates positive outcomes in well-conducted randomized controlled trials as gold standard evidence for predictions that the treatment in the trail will work if we implement it in our setting. So, what's the argument?

RCT results are normally *effect sizes: ES* = df the difference in the expectation of the outcome (y) in treatment group and in the control group $(\mathrm{Exp(y)_T} - \mathrm{Exp(y)_C})$. Causes do not, we suppose, produce their effects willy nilly, at least not where prediction is possible. Rather these effects are generated in accord with causal principles. We can without loss of generality suppose that these principles are of this form[1]:

$$\mathrm{CP: y(i) = a + b(i)x(i) + z(i)}$$

where *y*(*i*) is the outcome for individual *i* in the population where the principle holds, *x*(*i*) is the treatment variable, *a* is a constant and *z*(*i*) represents all the other casual clusters that contribute linearly with x to produce the value of y in i. It is apparent from this principle that x is a genuine contributor to y for at least some individuals i in this setting if $\mathrm{b(i)} \neq 0$ for some i. A well-known argument – which I shall call the *RCT Argument* – shows that, under usual assumptions about ideal RCTs,

$$\mathrm{ES = Exp\,(b)\,\left(X - X'\right)}$$

[1] The results I shall describe are essentially the same for more complicated functional forms.

where X = the value of the treatment variable in the treatment group and X′, the value in the control group.

RCT Argument

1. $y(i) = a + b(i)x(i) + z(i)$
2. $\text{ES} = \text{Exp}\,(y(i)/x(i) = X) - \text{Exp}\,(y(i)/x(i) = X')$

   $= \text{Exp}\,(a/x(i) = X) - \text{Exp}\,(a/x\,(i) = X')$

   $+\text{Exp}\,(b(i)/x(i) = X)X - \text{Exp}\,(b(i)/x(i) = X')\,X'$

   $+\text{Exp}\,(z(i)/x(i) = X) - \text{Exp}\,(z(i)/x(i) = X')$
3. x is probabilistically independent of b and w.

Therefore $\text{ES} = \text{Exp}(b(i))(X - X')$.

Premise (3) is supposed to be guaranteed by random assignment of individuals to the treatment and control groups and by masking, quadruple masking if possible. I shall suppose that it holds by definition in an *ideal RCT* and henceforth consider only ideal RCTs. We should remember of course that real RCTs are generally far from the ideal and that randomization only assures the independence assumptions in the long run were the same experiment repeated indefinitely.

So for an ideal RCT, if the effect size is positive, so is Exp(b) which means that b is positive for at least some i. So x is a genuine contributor to y for some individuals in a population subject to CP. This shows that there is a good argument, *A′*, that has among its premises the evidence claim

> e = df ‘The effect size of x for y in the population in a well-conducted RCT is ES > 0.’

and has as its conclusion

> h_1 = df ‘x is contributes to the production of y for some individuals in the population in that study.’

So e is evidence for h_1 *relative to* the RCT Argument and thereby relative to the other premises in that argument (including especially the assumption that conducting the experiment well – randomizing, masking, etc. –delivered the features an ideal RCT is supposed to have). To establish e’s evidential relevance to effectiveness prediction h, we now need to find an argument – a good argument – that I shall call the *Effectiveness Argument*, in which h_1 figures essentially as a premise and h as conclusion.

Before I propose one, I want to point out something about CP, which is often subject to a grave misunderstanding, one that I hope the reader won’t have been led into because I was careful with the notation. Often CP is written with the reference to the *i*’s implicit, so it looks like this:

$$\text{CP}' : y = a + bx + z.$$

In this case it is easy to suppose that b is a constant. But there are few treatment variables x for which this is likely be the case. After all, the treatment is usually only the salient factor, or the factor of focus, in a cluster of factors that together are sufficient to produce a contribution, that is, sufficient *when they all take the right values at once*. To use the terminology of JL Mackie (1965), x is a cause, yes; but it is an INUS cause of contributions to y: it contributes to y, but only when operating in cooperation with helping factors and often a great many of these. In CP, b(i) represents in one fell swoop the values for i of all the helping factors that are necessary along with x to ensure a contribution to y.

Now to the argument. First we need to formulate a conclusion properly. One version would be

> h_{ES} = df 'If x = X were introduced in our setting, as opposed to x = X′, keeping fixed all the other causes of y in our situation [except those downstream from x], the effect size would be ES for us too.'

So, will x make the same average contribution; that is, is the efficacy, which is measured by the treatment effect in the study situation, the same there as here. Certainly if the same principle holds there as here, *a* will be the same since it is constant. But *b* is not a constant; and the effect size is its expectation – that is, the effect size is an average over x's supporting factors. The average in each situation depends on the distribution of these in that situation. Even if the same principles govern the two, that is no reason to suppose the distributions of support factors would be the same. To the contrary in fact, this distribution very often heavily depends on local circumstances so it is unlikely to be the same.

Anyway, the same distribution is not really what you hope for. What you'd really like is that you have – or can arrange to have – a distribution that favours the good values of b – the ones that provide the largest contribution from the programme. At the least, you will want to have some values for which x's contribution is positive and these should outweigh the effects of those that make x's contribution negative; and if getting negative contributions in some individuals in your setting is to be avoided, then you don't want any of these at all.

Suppose though we can lay aside worries about negative contributions in some individuals. Suppose we want to predict simply

> h_{cont} = df 'If x = X were introduced in our setting, as opposed to x = X′, keeping fixed all the other causes of y in our setting [except those downstream from x], a positive contribution would result for some members of our population.'

What does it take to make ideal RCT evidence relevant? I am going to talk, for short, about whether x *can play a causal role* in the production of y – is it genuinely there in the principle for the production of y for some individuals? Here then is what I take it is the weakest valid argument that uses the results we can get from an RCT there as a premise and concludes that the programme or treatment will contribute positively for some individuals here.

Effectiveness Argument

1. x can play a causal role in the principles that govern y's production there.
2. x can play a causal role in the production of y here if it does so there.
3. The support factors necessary for x to make a positive contribution are present for at least some individuals here.

> Therefore, x can play a causal role in the production of y in some individuals here and the support factors necessary for x to make a positive contribution are present for at least some individuals here (i.e. x contributes to the production of y for some individuals here).

Where then does the RCT come in? It enters in a different argument, an argument that supports premise (1). That is why I talked earlier about what a study can evidence *directly*. As I use this term, a well-warranted empirical claim *e* is *direct evidence* for a hypothesis *h* if e figures essentially in a good argument for h – a valid argument with well-warranted premises. Now the RCT Argumnet is a valid argument that takes as premise a positive effect size in an experiment and as conclusion, that the programme contributes to the targeted outcome there in the study situation (post implementation). The other premises in the RCT Argument have to do with further features of the study; for instance that confounding factors are independent of x. The keepers of the evidence warehouses police these premises for particular studies: they judge how well-warranted the other premises in an argument like the RCT Argument are, mostly on the basis of the study design. So if we find a programme in a conscientious warehouse, we have good reason to think there is a good (valid and sound) argument like A′ to warrant the claim that x plays a causal role somewhere – there in the study setting. And that is the first premise in the Effectiveness Argument.

So the RCT result can be evidence for effectiveness here, but it is only *indirect*. It is not a premise in an argument for effectiveness but rather a premise in an argument for a premise. Moreover, its relevance is conditional, highly conditional, since it depends on the validity and the soundness of both the RCT and the Effectiveness Arguments. As in this picture, a positive effect size in an RCT is leveraged into evidence that the program works there (in the RCT setting) by the RCT Argument; and ‘it works there’ is leveraged into evidence for ‘it works here’ by the Effectiveness Argument; if either argument fails, the lever drops and evidential relevance disappears with a thud.

Both the RCT and the Effectiveness Arguments are valid, so what really matters is their soundness. We may take it for granted that the RCT Argument is pretty good if we find the programme in a reputable warehouse. What about the Effectiveness Argument? What ensures that its premises are well-warranted? Recall, the two additional premises necessary are:

2. x can play a causal role in the production of y here if it does so there.
3. The support factors necessary for x to make a positive contribution are present for at least some individuals here.

What further arguments support these premises? That’s the problem. There are no warehouses for information like this, and the kind of information needed is really hard to come by. I don’t see how (2) can be supported without a great deal of theory; so too with (3), in order to identify what the requisite support factors *are*. Then, in addition, (3) will require a good deal of local knowledge to determine if we have here even some of the right values for the support factors, let alone a desirable distribution of them.

Before returning to my overarching message, let me take up two objections to my account of what can count as warrant for an effectiveness prediction beyond the earlier objections to the Argument Theory in general.

First: RCTs are often advocated by people who don't like theory – they think our claims to theoretical knowledge are too slippery; they just don't want to trust to them. That means they don't like my view about how (2) gets warranted. They have an alternative proposal: more and more RCTs, with as much variation in circumstances as possible. I agree that more RCTs, and especially across a variety of circumstances, can improve the warrant for an effectiveness prediction. It does so by supporting a premise like (2): the program plays a causal role here. How? That's the rub. The argument could be by simple enumerative induction: swan 1 is white, swan 2 is white . . . ; x can play a causal role in situation 1, x can play a causal role in situation 2, . . . And how good is that argument? For induction we need not only a large and varied inductive base – lots of swans from lots of places; lots of RCTs from different populations. We also need reason to believe the observations are projectable, plus an account of the range across which they project. Electron charge is projectable everywhere – one good experiment is enough to generalise to all electrons; bird colour sometimes is; causality is dicey. Many causal connections depend on intimate, complex interactions among factors present so that no special role for the factor of interest can be prised out and projected to new situations.

I urge that rather than some weak inductive argument, we need a rigorous deductive argument. Then we know just what we are betting on when we bet on the conclusion. So I would add a premise to the effect that x can play the same causal role here as in all those other places, add it so that the challenge is clear: just what is the warrant for this very strong claim? That matters because of the weakest link principle: the conclusion can never have any more warrant than each of its premises individually.

The second objection is this. Surely the best evidence that the program will work here is an RCT here. I agree this would be good evidence – let's not quarrel about 'best'. *Would be* were it possible. But we never do an RCT here really, here on the same population at the same time. And both matter. A sample is almost never going to be a representative. Representative: that means governed by the same causal principles and having the same probability distribution over the causally relevant factors. And time certainly cannot be ignored. Are the causes the same now as they were when the study was done? That's a particularly pressing question for socioeconomic programme since economists from JS Mill to the distinguished British econometrician David Hendry have worried that past regularities are a poor guide to the future in economics, just because the background arrangement of causes shifts so often, and so unpredictably. Of course the experimental population could be representative enough and the causes at work stable enough. Let's just get this stated explicitly as one of our premises. Then we can think about what warrant there is for these assumptions in our case.

4 Conclusion

That returns us to my overarching point. Evidence is a 3-place relation; e is evidence for h only relative to some argument or other. That is not a new idea at all, and it may not be very controversial. But taking it seriously matters. It is altogether too easy, when we do not keep the arguments to the fore to overestimate the warrant that our studies can deliver. The RCT is a good example. It is widely taken in the evidence-based policy literature as gold standard evidence for effectiveness claims. Though perhaps with a caution. The US Department of Education, for example, warns that trials on white suburban populations do not constitute strong evidence for large inner city schools serving primarily minority students. This kind of warning simply conceals what needs to be exposed. What is the argument that makes a particular RCT result evidence for a particular effectiveness prediction? As we have seen, if evidence, it is indirect evidence – there are layers of arguments to get from the study result to the effectiveness conclusion. And they all have additional premises, every one of which, along the way, is essential for the security of the final conclusion. No matter how firm the RCT result is, the effectiveness conclusion – for which it is supposed to be gold standard evidence – can have no greater claim to knowledge than the shakiest of these.

Nor is this unusual. Most of our knowledge claims, even in our securest branches of science, rest on far more premises than we would like to imagine, and far shakier. This recommends a dramatic degree of epistemic modesty. Most of us have adjusted to Neurath's lesson that we are like sailors rebuilding our boat at sea. The conclusions I draw about evidence and the amount of warrant it can confer point to his less familiar warning: the boat is far leakier than we like to think.

Acknowledgement I want to thank Alex Marcellesi and the members of both the EPSA audience and the audience for my Pufendorf lectures who participated in the discussion for help with the immediate contents of this paper and the AHRC, the British Academy, LSE's Grantham Research Institute on Climate Change, the Spencer Foundation and the Templeton Foundation for support for the research for it.

References

Cartwright, N. (1999). *The Dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cartwright, N., & Fennell, D. (2009). Does Roush show evidence should be probable. *Synthese, 175*(3), 289–310.

Glymour, C., & Stalker, D. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly, 2*, 245–264.

Roush, S. (2005). *Tracking truth*. Oxford: Clarendon.

Models, Simulations, and Analogical Inference

Ilkka Niiniluoto

Abstract Models and simulations represent target systems by means of relations of similarity or analogy. Two objects or systems are similar if their attributes are close to each other or approximately equal. Two objects are analogous to each other if they are partly identical. From this perspective, it is useful to distinguish similarity models and analogy models as sources of learning about real targets. Similarity models include idealized models and their computer implementations which typically represent reality by deformation: while some irrelevant properties are excluded, some relevant properties are neglected by assigning them extreme values. Inferences from ideal similarity models are obtained either by approximation or by the concretization of counterfactual assumptions. Typical analogical models allow inference from the model to the target system by inductive inference from model data D to generalization C, and analogical reasoning from the model generalization C to the same generalization C about the real system.

1 Models as Representations

A shift of interest from theories to models can be observed in recent philosophy of science. Lively debates on the nature and function of models, with illustrations from natural, biological, cognitive, and social sciences, can be found e.g. in the special issues “Models and Simulations” (*Synthese* 169:3, 2009), “Economic Models as Credible Worlds or as Isolating Tools” (*Erkenntnis* 70:1, 2009), and “The Ontology of Scientific Models” (*Synthese* 172:2, 2010).

I. Niiniluoto (✉)
Department of Philosophy, History, Culture and Art Studies, University of Helsinki,
P.O. Box 24, 0001 Helsinki, Finland
e-mail: ilkka.niiniluoto@helsinki.fi

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_2,

The concept of *model* is often used for sets of special kinds of assumptions, including theoretical statements, mathematical equations, pictures, and diagrams. This notion is close to the traditional concept of *theory* or theoretical model. Following the Tarskian model theory, also set-theoretical *structures* or systems described by assumptions are called models. When a mathematical theory contains idealized assumptions, it describes *idealized models* which ignore or distort some aspects of real systems. Models may also be concrete artefacts. For example, *scale models* and miniatures are physical representations of prototypes. Today many models are *simulations* implemented by computer programs which allow for the systematic manipulation and variation of conditions.

Models are "epistemic artefacts" that can be used for various purposes (see Knuuttila 2005). A common feature of most of these cases is that models are used to *represent target systems*. According to the minimalist "inferential" account of representation (Suárez 2004), a model M allows competent and informed agents to draw specific inferences regarding its target R. Thus, it should be possible to ascertain facts about the target R by exploring, calculating, and experimenting upon the model M. For this purpose, some sort of relation of *similarity* or *analogy* should obtain between M and R. It is argued in this paper that this aspect of modeling has not been sufficiently developed in the recent discussions of models: in order to understand how a model M allows us to infer or learn something about the target R, one has to go beyond the minimalist account of representation and specify the similarity or analogy relation between M and R.

In the classical debate about analogical models, the "Duhemians" claimed that analogy has only a heuristic value in theory construction, while the "Campbellians" argued that analogical models are indispensable in science (see Hesse 1963; Hempel 1965). The latter view can be defended by noting that sometimes it is impossible to study an interesting target directly (Niiniluoto 1988). First, there may be computational limitations in our inferential capacities: the system R may be so complex that a simpler model M is needed; for example, equations can be simplified and approximated so that they have analytic solutions. Secondly, the current level of technology may restrict our possibilities of studying an inaccessible target R, so that a surrogate for R is needed. Thirdly, the study of the target R may be inhibited by moral reasons; for example, if R involves living human beings, its study can be replaced by animal experimentation.

For *heuristic* purposes, weak forms of analogy may be suggestive enough, if the discovered feature of M can then be tested by studying R itself. A famous example is Joseph Priestley's suggestion in 1767 that electric forces, which resemble gravitational forces, satisfy an inverse square law of attraction. Given the strong interest in the processes of discovery within artificial intelligence (see Holland et al. 1986), such applications of analogical inference are of high value. But analogy may play an important role in *justification* as well: in cases where the target R cannot or may not be directly investigated, our only way of justifying claims about R may be via the model M.

2 Inferences About the Target

One may distinguish three ways of obtaining scientific information about a research domain R of real entities in the world: experimentation, theorizing, and modeling.

Some research domains R can be directly *observed* by our senses and instruments or manipulated by *experiments*. When sufficiently many elements of R, which are of kind F, are observed to be of kind G, we may make an inductive inference to the generalization that all F's are G.

Classical forms of *theorizing* about R are based on linguistic descriptions D of R in some vocabulary L, where L may include both observational and theoretical terms. The target R may be described by statements in L, and such claims about R can be tested and confirmed by observational and experimental knowledge about R. Some aspects of R may be explained or predicted by theories formulated within the conceptual framework of L, and theories may be accepted "abductively" on the basis of their explanatory and predictive power.

According to scientific realism, the basic requirement for descriptions and explanations about R is *truth* (see Niiniluoto 1999). Critical realists acknowledge that even our best theories may be false, but still they may be truthlike: a theory in language L is *truthlike* with respect to the target R if it is close to the most informative true description of R in L (see Niiniluoto 1987). The weaker notion of *approximate truth* requires that the theory is close to some truth about R. Hence, truthlikeness means closeness to informative or comprehensive truth, and logically weak truths such as tautologies are approximately true but not truthlike. These concepts are applicable to theoretical models as well (see Niiniluoto 2002). By quantitative degrees of truthlikeness we can also define comparative truthlikeness (one theory is more truthlike than another) and dynamic truthlikeness (a sequence of theories converges to the complete truth about R). Relations of truthlikeness are reflected on the level of models of theories: for example, models of idealized theories are more or less close to the real system R, and "concretization" or the removal of idealized assumptions brings such models closer to R (see Nowak 1980; Niiniluoto 1990, 2007a).

In some cases, a theory T about the target R – even a true theory – may be too complex for our computational purposes, so that we are unable to derive a prediction about R from T. Then we may replace T with another theory T′ by *approximating* some claim in T (e.g. by giving the value zero or infinity to some constant or variable in the statements of T). T′ is then less truthlike than T, but we may be able to give an approximate derivation (explanation or prediction) of some empirical claim D about R from theory T′: D is *approximately derivable* from T′ if a statement D′ close to D is derivable from T′ (see Niiniluoto 1990).

Modeling is the third way of making inferences about target systems R. Typically such inferences rely on similar or shared features of a model M and a target R and project other features on this basis. The classical forms of this kind of inference are generalizations of the deductive rule for *identity*:

(RI) a is an F
b = a
Hence, b is an F.

Following Leibniz, the identity of a and b means that they share all of their properties F. Here a and b can be any real or abstract entities, individuals and structures.

A variant of RI is provided by the concept of *isomorphism* or structural identity: two structures A and B are isomorphic if there is a bijective mapping between the domains of A and B which correlates relations on A with their counterparts on B. This guarantees that there is a truth-preserving translation between the languages A and B so that every true statement about A is correlated with a true statement about B. In this way knowledge about a structure can be transferred to another isomorphic structure.

If the notion of identity in RI is replaced by likeness or resemblance, the rule is not deductively valid any more, but its strength depends on the "tightness" of the relation between a and b. This relation can be explicated in two ways (see Niiniluoto 1988, 2012).

First, according to the *similarity* interpretation, two objects are similar if their attributes are close to each other or approximately equal. A comparative notion "a is more similar to b than c" can be defined in qualitative terms, but more powerful quantitative degrees of similarity may be introduced if closeness is measured by numerical distances (e.g., 180 and 181 cm are close estimates) or distances in Carnapian quality spaces (e.g., red and orange are close colors). Suppose that a and b have w correlated attributes (such as height, color, etc.), and distances along each of these dimensions d_i, i = 1, . . . , w, are normalized, so that they take values between 0 and 1. Then the overall degree of similarity between a and b can be defined by means of the weighted Manhattan or Euclidean distance. As special cases of this general concept, we have the comparative and quantitative notions of structure-likeness (cf. Kuipers 2000). Inference by similarity can now be formulated by the rule

(RS) a is an F
b is similar to a
Hence, b is an F′,

where the features F and F′ are similar to each other. The strength of this argument depends on the degree of similarity between a and b. RS formulates the idea that similar causes bring about similar effects.

Secondly, according to the *analogy* interpretation, two objects resemble each other, if they are partly identical, i.e., they share many common attributes or relations (positive analogy) and disagree only on a few attributes or relations (negative analogy). This notion again has comparative and quantitative variants. Inference by analogy can be expressed by the rule

(RA) a is an F
b is analogous to a
Hence, b is an F.

In particular, the argument from positive analogy starts from the premises that a and b share several attributes $G_1, \ldots, G_{n,}$ and a is an F, and concludes that b is an F. Already J. S. Mill in 1843 suggested that the strength of RA depends on the relative sizes of positive and negative analogy between a and b, so that analogical reasoning is enumerative induction with respect to properties (rather than individuals).

More precisely, to avoid trivialization, the notion of analogy has to be defined relative to a conceptual framework with basic attributes (without their Boolean combinations). Objects a and b agree on attribute G if both of them or neither of them satisfies G. If a and b agree on k attributes and disagree on m attributes, their degree of analogy is definable by $k/(k + m)$. Other quantitative measures of analogy and similarity are defined in statistics, taxonomy, and psychology (see Niiniluoto 1987).

As defined above, similarity is a more general notion than analogy, as it allows that two objects are similar even when they disagree on all of their attributes (as long as the weighted combination of their distances is not too large). But analogy is a special case of similarity, where distances between attributes are replaced by the indicator function which has the value one when the attributes disagree and zero when they agree.

The distinction between similarity and analogy is relevant to the notion of truthlikeness as well (Niiniluoto 1987, 2002). According to the analogy interpretation, the truthlikeness of a theory presupposes that it has some matches with the complete true theory, or is *partially true*, but for the similarity approach this need not be the case. A theory as a whole may be truthlike even when all of its specific claims are false but close to the truth. The same conclusion holds for theoretical models: they can approximate a real system without being identical with it at any specific point.

Argument by analogy can be combined with inductive reasoning in many ways. For example, Kepler showed abductively, by carefully testing several alternatives, that the orbit of Mars is elliptical. He assumed that all planets (including the earth) are analogous with respect to their revolution around the sun, and concluded from one examined instance that all planets move around the sun along elliptical orbits.

Rule RA for singular analogy can be generalized to cases of *multiple analogy*, where the same conclusion receives support from several sources. This is relevant to formal accounts of inductive analogy. It turns out that the systems of inductive logic developed by Rudolf Carnap and Jaakko Hintikka satisfy the rule RA as a principle of positive analogy: observation of individuals of a certain kind increases the probabilistic expectation of finding further individuals of the same kind. But as soon as there is some negative analogy between the instances, this positive association breaks down. Systems of inductive analogy, where the probabilities are supplemented by an "analogy profit" from similar instances, have been developed as a solution to this problem (see Kuipers 1988; Niiniluoto 1988).

3 Analogy and Similarity Models

The conceptual tools of Sect. 2 give us a fresh perspective on modeling and simulation (see Niiniluoto 2012). It is useful to distinguish *analogy models* and *similarity models* as sources of learning about real targets.

Typical analogical models are real or concrete artefacts which allow inference from the model to the target system by rule RA. In a scale model (e.g., a map), the decreased dimension of the object is included in the negative analogy, but the relations between its parts in the positive analogy. Measurements of these relations can then be transferred to the target object. Another illustration is the use of animals as surrogates for human beings: experiments with mice and zebrafish can be treated as evidence about laws of human pathology and toxicology, since these animals have sufficient analogies with human physiology. On this basis, their reactions to some medical treatments can be transferred to human beings. In these cases, it is not true that experimentation deals directly with the target system (cf. Winsberg 2010, p. 52), as it is forbidden to do experiments with humans. Thus, the direct route from real data to a hopefully true generalization about human beings is not available, and it is replaced by the exploration of the model data D, inductive inference from model data D to generalization C, and analogical reasoning by RA from the model generalization C to the same generalization C about the real system.

Similarity models include idealized models which typically represent reality by deformation or caricature: while some irrelevant or less important properties are excluded, some relevant properties are neglected by assigning them extreme values. Ideal gas, as described by the Boyle-Mariotte law, is in this sense a model of real gas. Other examples are found in economics where models include unrealistic assumptions like perfect rationality and complete information (see Mäki 2009). Deductive and inductive inferences from the assumptions of such ideal similarity models would be at best truthlike or approximately true in the actual world, so that conclusions about the target should be based on the similarity rule RS. In other worlds, actually true conclusions would be at best approximately derivable from the ideal model. On the other hand, the model can be rewritten as a counterfactual conditional, which tells how the system would behave under the idealized circumstances. The idealizing assumptions are then the antecedents of the conditional, and this conditional may be true or truthlike (Niiniluoto 1990, 2007a; Hindricks 2012). To increase the accuracy of conclusions from the model, the counterfactual assumptions should be removed by "concretization" (Nowak 1980). For example, the ideal gas law $pV = RT$ is concretized by van der Waals law $(p + a/V^2)(V - b) = RT$, which introduces finite values to intermolecular attractive forces (a) and the size of gas molecules (b) (Niiniluoto 1999). van der Waals law then entails the counterfactual conditional $(a = 0 \ \& \ b = 0) \rightarrow pV = RT$.

According to Stephan Hartmann's (1996) definition, in simulation "one process imitates another process". Simulations resemble experiments in the sense that they can be "run" by a computer repeatedly. There may be a random element in the initial conditions (e.g., coin tossing) or in the steps of the process, so that these repetitions lead to different outcomes, but on the whole the process and its resulting pattern can be compared to some real phenomenon, such as climate change and economic behavior. Simulations may be implementations of ideal models. Some of them may have positive analogy with the target system, but typically they are similarity models. For example, Schelling's checkerboard model of cities exhibits a simple mechanism of segregation or racial sorting: two kinds of families occupy cells of a checkerboard, and when too many neighbors are alien, the family randomly moves to an unoccupied place. The members of these "toy cities" and their behavior do not share any attributes with real people in the sense of positive analogy, but still they bear some similarity with the intended target.

Robert Sugden (2002, 2009) has proposed that economic models are "credible counterfactual worlds", not abstractions or simplifications of reality, but rather fictional "parallel worlds" which are realistic in the same sense as novels. Sugden suggests that such credibility is the warrant for making inductive inferences from model to real world: in addition to real cities, imaginary or toy cities in Schelling's checkerboard model could serve as instances of inductive inference whose conclusion holds in the real world.

In my view, Sugden's idea of model-based induction should be modified and complemented by the notion of analogy (Niiniluoto 2012). We have already seen how an inductive generalization from a surrogate model is transferred to the real target by the rule RA of analogy. But this is against Sugden's fictionalism, since the analogical model (e.g., a toy city) should be then treated as a real and concrete artefact which by analogy gives some information about the target.

Another problem of Sudgen's account is that idealized economic models are not credible in his sense, as they include extreme assumptions. Such idealized models warrant counterfactual inductive inferences which lead to realistic conclusions by concretization or the similarity rule RS.

4 Successful Models and Realism

What should a scientific realist say about the fact that some false models are empirically successful? Many realists have argued that successful explanations and empirical predictions by a theory are good grounds for claiming that this theory is true or at least truthlike. This "no miracles argument" for realism can be based on Peirce's notion of abduction (see Peirce 1931–1935, 5.189), or inference to the best explanation, and it can be reconstructed in Bayesian terms as well: if theory H deductively or inductively explains evidence E, then E confirms H (see Niiniluoto 2007b). More precisely, the following formulation may be proposed:

(ES) If T consists of the postulates of a theory which are indispensable for the derivation of successful empirical consequences, then T is probably truthlike.
(See Niiniluoto 1999, p. 190, p. 193).

Eric Winsberg (2010) argues against the realist position that there are model-building techniques which do not purport to offer even approximately realistic or true accounts but still are empirically successful and reliable. His main example of such "fictions" is artificial viscosity, a special assumption of "an unphysical large value of viscosity", used by scientists working at Los Alamos.

There are several ways of replying to this kind of argument. Two of them are related to the notion of indispensability in ES, two to the relation between the premises and the conclusion in ES.

First, assumptions which are computationally indispensable, needed in practice for making calculations from a mathematical model, are simplifications and approximations, and taken in isolation need not be true or truthlike at all. Examples of such counterfactual extreme assumptions have been mentioned above. Classical mechanics is obtained from relativist mechanics by letting the velocity of light c to approach infinity, but the practical success of Newton's theory in engineering applications does not support the conclusion that c is infinite.

Secondly, when the ideal and extreme assumptions are incorporated into a larger context of more comprehensive principles, the theory or model as a whole may be truthlike to some degree in spite of the radical falsity of some the assumptions.

Thirdly, models which are only morally indispensable do not satisfy the conditions of ES. The success of animal experimentation in testing drugs does not show that men are mice.

Fourthly, while ES can be applied to theories and theoretical similarity models, concrete analogy models – even though they may epistemically warrant inferences about their targets - are not explanatory in the sense required by the abductive no miracles argument. ES can be understood so that the theory explains empirical consequences, and the explanans should be ontologically prior to the explanandum. This lack of explanatory ontological depth is illustrated again by animal experimentation. George Gamow's liquid drop model, which assumes the atomic nucleus to be a drop of incompressible fluid, was used successfully in the building of the atom bomb. The analogy between liquid and the nucleus is sufficient to develop Weizsäcker's formula which gives a useful approximation to the mass and binding energy of an atom, but atom bombs do not show that the atomic nucleus is a liquid drop.

Fifthly, as most scientific realists are fallibilists, it should be emphasized that the abductive support of theories by their empirical success is always uncertain and corrigible. Winsberg's formulation "success implies truth" (ibid., p. 121) is thus a misleading simplification. Many theories in the history of science have at first seemed to be highly successful, but later were refuted by new empirical evidence and replaced by other more truthlike theories.

References

Hartmann, S. (1996). The world as a process: Simulations in the natural and social sciences. In R. Hegselmann (Ed.), *Simulations and modeling in the social sciences from the philosophy of science point of view* (pp. 77–100). Dordrecht: Kluwer.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: The Free Press.

Hesse, M. (1963). *Models and analogies in science*. Notre Dame: The University of Notre Dame Press.

Hindricks, F. (2012). Saving truth for economics. In A. Lehtinen, J. Kuorikoski, & P. Ylikoski (Eds.), *Economics for real: Uskali Mäki and the place of truth in economics* (pp. 43–64). London: Routledge.

Holland, J., Holoyak, K., Nisbett, R., & Thagard, P. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA.: The MIT Press.

Knuuttila, T. (2005). Models, representation, and mediation. *Philosophy of Science, 72*, 1260–1271.

Kuipers, T. (1988). Inductive analogy by similarity and proximity. In D. H. Helman (Ed.), *Analogical reasoning: Perspectives of artificial intelligence, cognitive science, and philosophy* (pp. 299–313). Dordrecht: Kluwer.

Kuipers, T. (2000). *From instrumentalism to constructive realism*. Dordrecht: Kluwer.

Mäki, U. (Ed.). (2002). *Fact and fiction in economics: Models, realism, and social construction*. Cambridge: Cambridge University Press.

Mäki, U. (2009). MISSing the world: Models as isolations and credible surrogate systems. *Erkenntnis, 70*, 29–43.

Niiniluoto, I. (1987). *Truthlikeness*. Dordrecht: Reidel.

Niiniluoto, I. (1988). Analogy and similarity in scientific reasoning. In D. H. Helman (Ed.), *Analogical reasoning: Perspectives of artificial intelligence, cognitive science, and philosophy* (pp. 271–298). Dordrecht: Kluwer.

Niiniluoto, I. (1990). Theories, approximations, and idealizations. In J. Brzezinski et al. (Eds.), *Idealization I: General problems* (pp. 9–57). Amsterdam: Rodopi.

Niiniluoto, I. (1999). *Critical scientific realism*. Oxford: Oxford University Press.

Niiniluoto, I. (2002). Truthlikeness and economic theories. In U. Mäki (Ed.), *Fact and fiction in economics: Models, realism, and social construction* (pp. 214–228). Cambridge: Cambridge University Press.

Niiniluoto, I. (2007a). Idealization, counterfactuals, and truthlikeness. In J. Brzezinski et al. (Eds.), *The courage of doing philosophy: Essays presented to Leszek Nowak* (pp. 103–122). Amsterdam: Rodopi.

Niiniluoto, I. (2007b). Evaluation of theories. In T. Kuipers (Ed.), *Handbook of the philosophy of science: General philosophy of science – Focal issues* (pp. 175–217). Amsterdam: Elsevier.

Niiniluoto, I. (2012). The verisimilitude of economic models. In A. Lehtinen, J. Kuorikoski, & P. Ylikoski (Eds.), *Economics for real: Uskali Mäki and the place of truth in economics* (pp. 65–80). London: Routledge.

Nowak, L. (1980). *The structure of idealization*. Dordrecht: Reidel.

Peirce, C. S. (1931–35). *Collected papers 1–6*. Cambridge, MA: Harvard University Press.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71*, 767–779.

Sugden, R. (2002). Credible worlds: The status of theoretical models in economics. In U. Mäki (Ed.), *Fact and fiction in economics: Models, realism, and social construction* (pp. 107–136). Cambridge: Cambridge University Press.

Sugden, R. (2009). Credible worlds: Capacities and mechanisms. *Erkenntnis, 70*, 3–27.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago: The University of Chicago Press.

Intuitionistic Semantics for Fitch's Paradox

Doukas Kapantaïs

Abstract If one, in order to evaluate ¬Kp, follows the BHK condition for negated formulas, and takes Kp to be untrue in all possible worlds, Fitch's paradox is no threat to the antirealist. However, the semantics become intolerably inexpressive. On the other hand, if one interprets ¬Kp as saying that Kp is untrue *in the actual world*, another way-out of the paradox presents itself. I sketch which one this is, and I describe the intuitionistic models, in which it can be applied. I show that, within these models, one can built the knowability principle, while, at the same time, not everything is known in every world. Moreover, by applying the Beth condition for existential quantification, I show that there are worlds of these models, where a sentence saying that we will come to know something we now ignore is true/established. So, in these models, where the knowability principle holds good, not only are there worlds, in which not everything is known, but these worlds can prove that much as well.

1 Strong Negation and Fitch's Proof

Fitch's paradox is built upon two premises and a limited number of some quite inoffensive modal and deductive rules. The first premise says that every truth can be known, the second, that there is at least one unknown truth. Now, say that I am an antirealist and also (this is frequent in case one is an antirealist), an intuitionist with respect to logic. Intuitionist I might be, but no fool am I, and so I do not believe that every single truth is known. However, I do believe (at the end of the day this is what my "antirealism" amounts to), that all truths can be known. In that respect, I am the perfect victim for one to play the Fitch paradox on.

D. Kapantaïs (✉)
Academy of Athens, Research Centre for Greek Philosophy, Anagnostopoulou 14,
106 73 Athens, Greece
e-mail: dkapa@academyofathens.gr

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_3,

But wait a minute, what have I just said? I have introduced myself as an intuitionist with respect to logic, and I have added (in an effort not to sound silly), that not all truths are known. And now the question arises: how come and I, an intuitionist with respect to logic, happen to know there actually being one such truth? There has to be one, I am sure (this is the "I do not want to sound silly" part), but how I, the intuitionist, am I allowed to formally claim that much? As a matter of fact, and according to some assumptions not at all foreign to my intuitionism, and more precisely, according to the paradigm interpretation of negation in intuitionism, not only am I not allowed to claim that much, but, moreover, I cannot even articulate a formula expressing what I am not allowed to claim. In the following paragraph, I explain why.

Intuitionistically, p being true amounts to having a proof of p, which, in case p is a negative statement, amounts to being in position to transform any proof of what p negates into a proof of a contradiction. Let p be a sentence that is true and unknown. If no one knows p, no one has a proof of p either. Now, it must also be the case that, since p is true, no one can have any proof that p can be transformed into a contradiction either. Let us apply the same truth criterion to "p is unknown". Since "p is unknown" is a negative statement, we need to have a proof that the assumption that p is known leads to contradiction. Again, this is exactly what we cannot have in case p is true, and the knowability principle holds good. For, if the knowability principle holds good, there must be a possible world, where someone knows p. But I, the antirealist, wish exactly that: that the knowability principle holds good. And so, in case p is true, I am very much content with the idea that "p is unknown" is provably false. For my "p is unknown" does not say that no one knows p in the actual world, but that the assumption there being one such person is contradictory, which amounts to saying that "p is known" is untrue in every possible world, and this is exactly what I do *not* want, in case p is true and I believe in the knowability principle.

All this follows from an apparently innocent wish of mine to conform to the rules revealing the truth conditions for ¬p in intuitionism, according to the paradigm interpretation of intuitionistic negation. What can be wrong with that? *Prima facie* nothing at all. And I have an extra bonus too. I cannot possibly prove any formula saying that there are no unknown truths (i.e. no one can accuse me of omniscience). For, in order to prove such a formula, I must first be in position to write down another formula saying that there is one unknown truth, and I cannot do that. My $(\exists p)(p\&\neg Kp)$ says what "there is an unknowable truth" says in English.

2 The Shortcomings of this Solution

An antirealist – we have seen – if she also happens to be an intuitionist with respect to logic, has an easy way-out of the paradox. In fact, the "paradox" becomes for her a proof of some formulation of the knowability principle. From the premises: (i) all truths are knowable and (ii) there is an unknown (=unknowable) truth, a contradiction can be derived. Hence, our antirealist readily rejects (ii).

The joy of the above antirealist will be short-lived. For the realist straightforwardly strikes back by wondering what kind of logic/semantics is this (i.e. the intuitionistic) that not only can it not capture the self-evident truth that there are unknown truths, but, moreover, it cannot even articulate any statement claiming that much. "Unknown" cannot be distinguished from "unknowable".

It is not only this. Another distinction, most essential for her argumentation against classical logic, becomes blurred as well. For the same reason that she can no longer distinguish between unknowable and unknown, she also misses the distinction between p and "p is known". As a matter of fact, these two extensionally collapse, and one can no longer formally describe the specific kind of situation, where a sentence is truth-valueless and, therefore, unknown. For, again, if one follows the paradigm intuitionistic criterion for negation, and p is truth-valueless, then, Kp and ¬Kp are also truth-valueless. In some more detail, this happens for the following reason. If p is truth-valueless, there has to be a possible world where p is proved and some possible world where it is not. From which, it follows that there has to be a possible world where p is known, and another, where it is not. Hence, "p is known" is also truth-valueless. This is quite harmful with respect to the intuitionistic expressive resources. Consider, e.g., an actually truth-valueless – from the point of view of the intuitionist – sentence: the formula stating Goldbach's Conjecture. This formula, by being truth-valueless, is incompatible with anyone actually knowing what it says, and anyone actually knowing what its negation says. But what does this imply for the *Weltanschauung* of our intuitionist? It implies that, for her (i.e. *she* also claims that), we do not actually know either the formula stating the Conjecture or its negation. However, as it so happens, this evident truth (evident in the eyes of the intuitionist as well as the realist) cannot be captured by these semantics.

One thing that can be done in order to remedy these shortcomings is to adopt a decidable metalanguage with respect to "p is known", at least for cases, where p is recognized as truth-valueless; i.e. we minimally expect the intuitionist to be in position to claim that she does not know p, in case she takes p to be truth-valueless. By doing this, we (i) extensionally dissociate p from "p is known", for "p is known" is false, when p is truth-valueless, and (ii) we can distinguish between unknown and unknowable, for when p is truth-valueless, p is both unknown and not unknowable. It is unknown, because p is untrue in the actual world, and it is not unknowable, because, since p is truth-valueless, there has to be at least one possible world, where p is known. As a matter of fact, and as will be made clear in the following section, by making "p is known" decidable, we make the negation in "p is unknown" behave exactly like weak negation. "p is unknown" is interpreted as "p is not known now" and not as "p cannot be known".

3 A Dilemma for the Intuitionist

That p is false necessitates not only that p is untrue, but, moreover, that p is untrue in every possible world, which is the same as saying that from the mere assumption of p – from this assumption alone – a contradiction can be derived. For "p is known" to be false what is necessitated is not that "p is known" is untrue in every possible world but, rather, that "p is known" is untrue in the actual world. If we add to the above that one has always sufficient evidence as for whether "p is known" is untrue in the actual world, one makes "p is known" decidable. Notice that the general intuitionistic criterion for assigning a truth-value is not thereby abolished with respect to "p is known". "p is known" is recognized as true-or-false not independently of any evidence one has about its truth/falsity. On the contrary, it is recognized as such, because it is recognized as false, and it recognized as false, because some contradiction can be derived from the assumption that it obtains. But, this time around, the contradiction does not emanate from the mere assumption that p is known, but from the assumption that p is known *in the actual world*. Consequently, while "p is unknown" becomes true when "p is known" is untrue, we still keep a unitary interpretation for negation: i.e. as saying that what it negates leads to contradiction.

Let us now turn our attention back to Fitch's proof. Assume that the Excluded Middle is reestablished for "p is known", and for the reasons stated above. Or, minimally, assume that, when p is recognized as truth-valueless, "p is known" is recognized as false.[1] "p is unknown" can now be distinguished from "p is unknowable". But what does this imply for the intuitionist battling her way out of the paradox? It implies this. The all too easy solution the intuitionist has had up to now is no longer available. For, as soon as "p is unknown" stops being equivalent with "p is unknowable", the second premise of Fitch's proof (i.e. the one saying that there is one unknown truth), stops being the negation of some formulation of the knowability principle, and, so, Fitch's proof, ending with the refutation of that premise, is no longer something the antirealist should be proud of; in fact, it acquires anew its old paradoxical flavor. It says, again, that everything is known.

Appearances are that the realist has dragged her opponent into a dialectical impasse. The latter can either make "p is unknown" of the second premise of Fitch's paradox as undecidable as p and suffer the consequences of Sect. 2, or she can make the same "p is unknown" decidable and suffer the consequences of Fitch's paradox: actual omniscience.

In what follows, I argue that this is a false dilemma, since the intuitionist can, after *and because* of these developments, follow another path out of the paradox. The "... and because of these developments" clause above means: what the intuitionist can now claim against the validity of Fitch's proof is not independent from "p is known" having been made decidable, and from the reasons it has been made decidable. The solution bears resemblances to the solutions already proposed in Edgington (1985, 2010), and Kvanvig (1995, 2006), but most importantly in Brogaard and Salerno (2007).

[1] A moment's thought will convince you that the crucial question here is whether the intuitionist can recognize something as truth-valueless. By the moment that she does, one has positive evidence that she can impose decidability on Kp as well.

4 Weak Negation and Beth Condition for Existential Quantification

The spirit, if not the technicalities, of this solution is the following. We have taken the truth-value of "p is known" to be a function of the possible world, in which this claim is made and of this world only. The same, of course, has to be the case with respect to "p is unknown". So, when one claims that there is a truth, which is unknown, one does not thereby claim that there is a truth that is unknown in all possible worlds, or in some of them, as for that matter, but a truth that is unknown in the actual world. Suppose that this truth is represented by p. The combined claim generating Fitch's paradox is "p and p is unknown". By the knowability principle, it follows that the truth represented by the above sentence is possibly known. Consequently, there is a possible world, where "p and p is unknown" of the actual world is known to be true. Therefore, p is known to be true in this possible world, and "p is unknown" is known to be true in the same possible world. But the initial meaning of "p is unknown" was not involving the claim that p is unknown in this other possible world, where "p and p is unknown" is known to be true; the initial claim was that p is unknown in the former. As a matter of fact, that these are (need to be) two different worlds is what makes the paradox disappear. For what is known in the possible world, where "p and p is unknown" is known to be true, is (i) that p and (ii) that p is unknown in this other possible world, where it is unknown; i.e. not in the world, where "p and p is unknown" is known to be true. This means that when we distribute the knowledge operator over "p and p is unknown", no contradiction can emerge. For after the distribution, we end up with two different non contradictory claims. We end up with the claim that (i) p is known in the possible world, where "p and p is unknown" is known to be true, and (ii) in the same world, p is known to be unknown in the former world, i.e. in the world where it is unknown indeed. From (ii) we can, of course, deduce that, in the world where p is unknown, p is unknown, but there is not any problem with such a conclusion.

The point is as simple as that. Denying it would be as strange as claiming that it is contradictory for me to claim at present that I do not know the truth of "it rains in Paris and I do not know that it rains in Paris", and that, if tomorrow I arrive at knowing it, part of what I will know then will be the truth of "I do not know that it rains in Paris"! When I say that I do not know p, I mean that I do not know p at present. And when I say that I do not know p at present, I have some moment of today in mind, not what "at present" will be denoting, e.g., sometime tomorrow.

The intuitionistic models where this solution can be put into work have been invented by the author (Kapantaïs 2009, Chap. 4, 2011) they are conservative extensions of Beth models for intuitionistic logic (Beth 1956; Van Dalen 1986, pp. 249–252), and have an integrated metalanguage. I have named them "S-models", S standing for "*Scientist*": a *Creating Subject* dealing not only with mathematics. I will briefly present the Fitch-relevant parameters of some simplified version of them in the following section. These parameters are more ambitious than the mere blockage of Fitch's proof. In reality, it is not only that the intuitionistically valid conclusion of the proof (i.e. ¬(∃p)(p&¬Kp)) is shown not to be a threat, but, within the same models, there are worlds where a sentence saying that there is a truth that will be known, but is actually unknown, is not only true/proved *about* them, but also true *within* them. This is arguably a step forward, since it allows the intuitionist to assert that there is something she does not actually know, even if she has not constructed any witness for this "something" as yet.

To be precise, intuitionists, now and again, have allowed themselves such conclusions, even without the presence of any witness. They have allowed themselves these conclusions on the bare evidence of the unavoidability of constructing one such; i.e. on the actual evidence that such a witness will, sooner or later, be constructed (see, for example, Brouwer 1981, p. 92; Dummett 1977, p. 6).[2] The more liberal approach allowing witnessless proofs is formalized by Beth models; the rigid one, which forbits them, is formalized by Kripke models for intuitionistic logic (Kripke 1965). Intuitionistic semantics can apply this particular solution of the paradox uniquely within Beth-like models.

5 Formal Account

Consider an infinite set of worlds W, put into a partial order by an accessibility relation R, such that W becomes an infinite in both directions, finitely branching in the forward direction, tree.

Intuitive background: the actual present is one of these worlds. Its linear past lays behind; its branching future, ahead. Each world sees itself and all the worlds coming after it within the tree it defines. R captures the "... is a possible present or future universal state of---" relation.

Assume a function from the Power Set of the set of atomic sentences to the elements of W. Take the closure of sentences made by atomic sentences and ¬, &, ∨, →, (,), under wellformedness.[3] The truth-value of sentences that belong to this closure is determined by the intuitionistic Beth conditions for the connectives, as put forward in, e.g., (Van Dalen 1986, pp. 249–250, p. 264).

Intuitive background: These sentences represent things that obtain in worlds but contain no reference to any world. (e.g. "It rains" belongs to this closure; "It will rain tomorrow" does not.)

We introduce references to worlds from within the worlds of the model themselves, by internalizing the metalanguage. Consider the usual "forces" relation of intuitionistic semantics. "... forces ($\Vdash$)---" relates worlds to sentences in the sense that a world "forces" a sentence just in case the sentence is established/true (one has evidence for what the sentence says) in that world. We internalize it as follows:

[2] Thanks to an anonymous referee for this point.

[3] We skip the predicate logic part here, since it is inessential for the argument.

If p is well formed and w a world, $w \Vdash p$ is well formed. (NB: “..is well formed” in the internalized sense, i.e. in the sense that it belongs to the kind of items that belong to the worlds.)

Now we pose:

(i) w forces $(w' \Vdash p)$ if and only if w′ forces p.
(ii) if w does not force $(w' \Vdash p)$, w forces $\neg(w' \Vdash p)$.

Intuitive background: (i) and (ii) capture the intuition that, whereas something being the case might be non decidable in some possible universal states (possibly the actual one), whether or not something is established to be the case, is always decidable in every possible universal state and with respect to every possible universal state. To better see what lays behind this principle, think of an intuitionistically idealized scientist, who might be ignorant of whether something obtains or not, because there is no state of affairs accounting either for the sentence representing it or for its negation, but is never ignorant of whether something is intuitionistically established in any world w. (Notice that the scientist might know that p is not established in w and not know that the negation of p is established in w. If so, the scientist knows also that the negation of p is not established in w.)

One can now introduce tenses other than the present via the internalized metalanguage as follows. Here, we will exhibit the future tense only:

Two definitions are needed. A “path through w” is a maximal linear order passing through w. A “bar for w” is a set of worlds containing a world from every path through w.

(iii) w forces F(p) if and only if there is a bar *B* for w, such that, for every world $w' \in B$, $w' > w$, and w′ forces $w' \Vdash p$.

Intuitive background: That something is the case necessitates intuitionistically that we have evidence that it is the case. That something will be the case necessitates that we have actual evidence that it will be the case (in the classical idiolect one would have to add “… no matter what”). Hence, the bar condition in (iii): p will happen if and only if p happens sometime in *every* possible future.

We now internalize the “… is known” predicate by K(…), and as follows:

(iv) w forces K(p) if and only if w forces p.
(v) if w does not force K(p), w forces ¬K(p).

The intuitionistic *rationale* behind (iv) should be obvious. Only what is established is known and vice versa. (v) imposes the decidability of K(…).

A comment on our implied *Signature* is necessary at this point. The *Signature* of S-models is of the same nature as the *Signature* of Beth models, but it is augmented by (among other things) a depository of names for sentences. What is important here is that the fundamental function of names does not change in between names for individuals and names for sentences. They both help us to identify things throughout and also *in* possible worlds. Now, if we wish to be in position to identify same propositions throughout possible worlds, we need sentences picking up same propositions in different worlds. If the stipulations in Sects. 3 and 4 are followed,

$\neg$Kp does not fulfill that purpose. For, although $\neg$Kp of w expresses the proposition that p is unknown in w, $\neg$Kp of $w' \neq w$ expresses the proposition that p is unknown in w'. On the other hand, for any mathematical p, p expresses the same proposition in every possible world. We will say that a sentence is "eternal" if and only if it expresses the same proposition in all worlds. So, if we wish to identify same propositions in different worlds, we need eternal sentences. In particular, and with respect to Fitch's proof, we need of some sentences expressing in all worlds the proposition that $\neg$Kp expresses in the actual world. We construct these kind of sentences as follows.

First, observe that, because of the internalization of the metalanguage, we have a simple way to express in any world w the claim that p is forced in some world w' (not necessarily different from w); we can use $w' \Vdash$p.

We now pose:

(vi) If p belongs to w, w $\Vdash$ p is its "eternalized counterpart".

This does the trick for $\neg$Kp. Observe that, unlike $\neg$Kp, which, according to the world it belongs to, expresses the proposition that p is unknown in that world, its eternalized counterpart expresses the proposition that p is established to be unknown in some constant world, and not in the world it occasionally belongs to. Therefore, if we need to identify in other possible worlds the claim made by $\neg$Kp in w we just use its eternalized counterpart.

We are now in position to give a sketch of the formal proof of the first of the two claims made at the end of Sect. 4.

We begin by imposing the knowability principle to the models as follows:

KP: If something is true, there is a world of the model where it is known.[4]

One would perhaps find this gloss of the principle a bit too generous. For, within its initial formulation, there is this world "possibly" (all truths are *possibly* known), and the tone for the possibility operator is usually given by the accessibility relation, which, in S-models, is confined to worlds that can be – so to speak – "visited" from the actual world (possible futures) or are the actual present. This captures the intuition that the past cannot happen again. I do not take it, however, that the knowability principle says that all truths can be known in some possible *future*. The antirealist/verificationist has made life difficult for herself right away after having endorsed the knowability principle; i.e. right from the start. She really does not need to make it practically impossible. What the verificationist minimally needs is some argument defending the view that understanding a sentence amounts to understanding the conditions under which the sentence is established and known to be true. In that respect, *any* world of the frame would do; what happens in any

[4] This formulation belongs to our external, classical, metalanguage, but concerns eternal sentences of the *Signature*. This is important to notice, since it might be the case that one such sentence is true *sub specie* of our external metalanguage, while, at the same time, it might be untrue in some worlds of the model. I do this for simplicity. The principle can be built with equivalence classes of "being *now* true" sentences.

world of the frame is – according to an armchair philosophical attitude of course – "understandable", and no matter whether one, *anyone*, will ever live to experience it or not. So, the formal interpretation of "possibly" with respect to the knowability principle should – I think – be as generous as it can.

Now, assume that p is true, eternal and unknown-in-w.[5] Eternalize $\neg K(p)$ of w by $w \Vdash \neg K(p)$, in order to be able to capture the claim involved in $\neg K(p)$ of w, across possible worlds. (p, since eternal, already identifies the same proposition across worlds.) By the knowability principle, it follows that there is a world w', where $(p \& w \Vdash \neg K(p))$ is known. So, $K(p \& w \Vdash \neg K(p))$ is true in w'. From which, it follows that K(p) is true in w' and $K(w \Vdash \neg K(p))$ is true in w'. No contradiction can be derived from there. In order to see this better, eternalize K(p) of w' by $w' \Vdash K(p)$, and apply the factivity of K in order to obtain $w \Vdash \neg K(p)$. Both $w \Vdash \neg K(p)$ and $w' \Vdash K(p)$ are established in w', which is exactly as required: In w', p is established to be known in w' and it is also established not to be known in w.

We have proved that there are S-models, where **KP** is valid and some truths are unknown in some of their worlds. (The trick is, again, that $\neg(\exists p)(p \& \neg Kp)$ is valid in the models, but what it says is that there is no sentence that is *in the same world* established and unknown.)

By the Beth condition for intuitionistic existential quantification (but not with the Kripke condition) we can do even better. We can show that, in some worlds of the model, people know that they will come to know some truths they actually ignore; i.e. we can prove the second claim at the end of Sect. 4.

In order to do this, we need to apply the Beth condition for existential quantification for sentences falling under the F(. . .) schema:

(vii) w forces $(\exists p)F(\varphi(p))$ if and only if there is a bar *B* for w, such that for every world $w' \in B$, (i) $w' > w$, and (ii) there is a sentence q, such that $\varphi(q)$ is forced in w'.

Notice: (1) φ is a metavariable belonging to our external metalanguage; a variable of sentential predicate expressions of the *Signature*. (2) q is under the scope of "for every world w'", so it need not be the same sentence in every world of the bar. In substance, this is what the Beth condition amounts to.

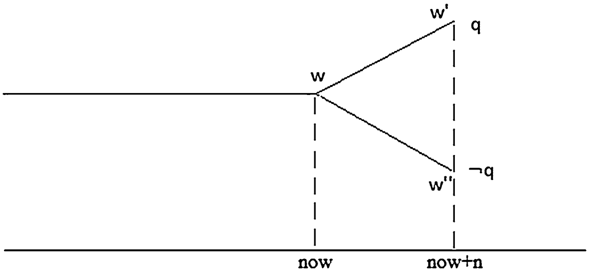
Intuitive background: If we have evidence that, in all possible futures, some sentence (it might not be the same), will have some property, we can already assert that *some* sentence will have this property. The bar represents this evidence.[6]

Now, assume that we know (by, e.g., Complexity Theory) that a Machine executing an algorithm can decide the open mathematical problem Q, in exactly

[5] Notice, again, that this sentence makes part of our external metalanguage. This is important, because p is *not* true in the worlds, where it is unknown, as neither is $p \& \neg K(p)$.

[6] Notice that the more general schema: "w forces $(\exists x)f(x)$ if and only if there is a bar *B* for w, such that for every world $w' \in B$, (i) $w' > w$, and (ii) there is a y, such that f(y) is forced in w'", should not be valid. For example consider the case where sometime in every possible future there lives a person more than 200 years old. The non validity of the more general schema is the reason why $\neg(\exists p)(p \& \neg Kp)$ is valid in the models. It does not depend on future bars but is evaluated in situ.

n moments from now, and that we also have some independent evidence that the Machine will not fail doing so (e.g. because of some physical hindrance). Say that the actual world is w. Assume that the sentence giving a positive answer to the problem is q. Since Q is mathematical, $q/\neg q$ are eternal. By the decidability of $K(\ldots)$ (see (v)), it follows that $\neg K(q)$ is established in w, and that $\neg K(\neg q)$ is established in w. Eternalize $\neg K(q)$ and $\neg K(\neg q)$ of w, by $w \Vdash \neg K(q)$ and $w \Vdash \neg K(\neg q)$ respectively, and for the reasons mentioned in the previous proof. By (i), it follows that they both belong to w. Now, say that in the world w', such that $wR^{n}w'$, the Machine gives a positive answer to Q, and in the world w'', such that $wR^{n}w''$, the Machine gives a negative answer to Q.[7] (See figure.) By (i), it follows that both $w \Vdash \neg K(q)$ and $w \Vdash \neg K(\neg q)$ belong to w' and w''. By (iv), it follows that w' forces $K(q)$ and w'' forces $K(\neg q)$. From which, it follows that $(K(q))\&w \Vdash \neg K(q)$ belongs to w' and $(K(\neg q))\&w \Vdash \neg K(\neg q)$ belongs to w'' (intuitionistic conjunction functions the same as classical conjunction). Notice that $\{w', w''\}$ is a bar for w. By applying (vii), $(\exists p)F(K(p)\&w \Vdash \neg K(p))$ immediately follows in w, i.e. take $(K(\ldots))\&w \Vdash \neg K(\ldots)$ for φ.[8]



6 Conclusion

First, we have shown that, if one interprets $\neg Kp$ by some (arguably the paradigmatic) intuitionistic interpretation of negation, Fitch's paradox can be blocked. We have then pointed out that this interpretation of $\neg Kp$ runs into serious difficulties having to do with the expressive resources of the underlying logic. By abandoning it, and making $\neg Kp$ decidable, another line of approaching the problem has emerged. This is to say that, by showing that the correct intuitionistic reading of $\neg Kp$ is "p

[7] If you do not feel at ease with both q and ¬q being possible, when q is mathematical, consider the example in the following note.

[8] A more everyday life candidate for q is the following: Think of any decision you have not made as yet, that you can, in principle, make, and which is still within your powers either to make or not to make; then, construct an eternalized sentence for today's "I will have taken this decision by the end of the day".

is not known in the actual world", we have shown that a solution to the paradox along the lines of the family of solutions mentioned at the end of Sect. 3 not only is available to the intuitionist but it is the appropriate one too. Formally, we have presented this solution in some models of temporal intuitionistic logic with internalized metalanguage and with the knowability principle built into them. We have – in an initial stage – shown that in these models the knowability principle holds good, while, at the same time, not everything is known in every world. In a second stage, and by adopting the Beth criterion for existential quantification, we have proved that in some of their worlds a sentence, saying that some truth that is now unknown will be known in the future, becomes proved.

References

Beth, E. (1956). Semantic construction of intuitionistic logic. Kon. Nederlandse Ac. Wetenschappen afd. *Letteren: Mededelingen, 19*(11), 357–388.

Brogaard, B., & Salerno, J. (2007). Knowability, possibility and paradox. In D. Pritchard & V. Hendricks (Eds.), *New waves in epistemology* (pp. 270–299). New York: Palgrave Macmillan.

Brouwer, L. E. J. (1981). In D. van Dalen (Ed.), *Cambridge lectures on intuitionism* (Appendix). Cambridge: Cambridge University Press.

Dummett, M. (1977). *Elements of intuitionism*. Oxford: Oxford University Press.

Edgington, D. (1985). The paradox of knowability. *Mind, 94*, 557–568.

Edgington, D. (2010). Possible knowledge of unknown truth. *Synthese, 173*(1), 41–52.

Kapantaïs, D. (2009). *The sea-battle and intuitionism*. Revised and augmented version of my Ph.d. Thesis, University of Bern (2007).

Kapantaïs, D. (2011). Intuitionistic formal semantics for future contingents. In P. Nomikos (Ed.), *Proceedings of the eighth panhellenic logic colloquium, Ioannina* (pp. 68–72). University of Ioannina.

Kripke, S. (1965) Semantical analysis of intuitionistic logic I. In M. Dummett & J. Crossley (Eds.) *Formal systems and recursive functions*. *Proceedings of the Eighth Logic Colloquium, Oxford* (pp. 92–130). Amsterdam: North-Holland.

Kvanvig, J. (1995). The knowability paradox and the prospects of anti-realism. *Noûs, 29*, 481–499.

Kvanvig, J. (2006). *The knowability paradox*. Oxford: OUP.

Van Dalen, D. (1986). Intuitionistic logic. In D. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (Vol. III, pp. 225–339). Dordrecht: Reidel.

Correlation and Truth

Peter Brössel

Abstract The concept of correlation is the building block of almost any Bayesian attempt to capture or explicate any interesting aspect of scientific reasoning in terms of probabilities. This paper discusses one particularly simple correlation measure which is highly significant for almost any such attempt within the philosophy of science or epistemology. In particular, it shows how this correlation measure is related to central attempts to capture essential aspects of scientific reasoning such as confirmation, coherence, and the explanatory power of hypotheses. This intimate connection between correlation and scientific reasoning necessitates answering the question of how correlation and truth are related. This paper proposes an answer to this question and outlines its consequences for epistemology and the philosophy of science.

1 Introduction

The qualitative concept of correlation is easily understood. Two propositions A and B are correlated relative to a probability function Pr if and only if $\Pr(A \cap B) \neq \Pr(A) \times \Pr(B)$; alternatively if and only if $\Pr(B) > 0$ and $\Pr(A|B) \neq \Pr(A)$.

The simple concept of correlation is the starting point of almost any Bayesian attempt to capture or explicate any interesting aspect of scientific reasoning in terms of probabilities. The following examples illustrate this point clearly. The central aim of Bayesian confirmation theory is to provide an explication of the qualitative ("evidence E confirms hypothesis H with respect to background knowledge B") and the quantitative notion of confirmation ("evidence E confirms hypothesis H with respect to background knowledge B to degree r"). Intuitively, the degree of

P. Brössel (✉)
Department of Philosophy, University of Mainz, 55099 Mainz, Germany
e-mail: broessep@uni-mainz.de

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_4,

confirmation of a hypothesis in the light of the evidence should reflect how much the hypothesis is supported by the evidence and how worthy of belief it is given the evidence. According to the Bayesian standard conception of the qualitative notion of confirmation, evidence E confirms a hypothesis H (with respect to background knowledge B) if and only if E and H are positively correlated (in the light of background knowledge B).

One can find many more examples like these in Bayesian philosophy of science and epistemology, and some of them will be discussed in the following section. What all these examples demonstrate is this: correlation is intimately related to exactly those aspects of scientific reasoning that Bayesian philosophers of science and epistemologists seek to understand. This paper focuses on the study of correlation. Section 2 presents one particularly simple correlation measure which is highly significant for the philosophy of science and epistemology. In addition, Sect. 2 demonstrates how this correlation measure is related to important aspects of scientific reasoning: confirmation, coherence, and the explanatory power of hypotheses. The intimate connection between correlation and scientific reasoning evokes the question of how correlation and truth are related and what possible consequences this relation has for those aspects of scientific reasoning that are so closely linked to correlation. This question is answered in Sect. 3. Section 4 outlines the consequences the presented results have for the philosophy of science and epistemology from a Bayesian point of view.

2 The Role of Correlation in Bayesian Epistemology

2.1 Correlation Measures

In the philosophy of science and epistemology, interest typically centers on whether certain propositions correlate. Thus, the typical definition of the degree of correlation between random variables which can assume more than two values is not applicable. Given this need for a measure of correlation between propositions, a particularly simple correlation measure can be shown to be fruitful for the Bayesian philosophy of science and epistemology.

Definition 1 (Simple Correlation).[1]

$$\mathfrak{corr}(A_1, \ldots, A_n) = \frac{\Pr(A_1 \cap \ldots \cap A_n)}{\Pr(A_1) \times \ldots \times \Pr(A_n)}$$

if $\Pr(A_i) > 0$ for all $1 \leq i \leq n$, and 0 otherwise.

[1] Wheeler (2009) calls this the Wayne-Shogenji correlation measure. Wayne (1995) discusses whether $\mathfrak{corr}$ can be taken to be a similarity measure and Shogenji (1999) interprets it as

2.2 Correlation and Confirmation

Confirmation theory is one of the central fields of application of the Bayesian machinery and it is intimately related to correlation. As already indicated at the outset, Bayesian confirmation theory holds that some evidence E (incrementally) confirms a hypothesis H relative to (background knowledge B and) probability function Pr if and only if $\Pr(H|E) > \Pr(H)$ (respectively $\Pr(H|E \cap B) > \Pr(H|B)$). Hence, E (incrementally) confirms H just in case E and H are positively correlated (in the light of the background knowledge B) relative to probability measure Pr (in what follows, mention of the background knowledge B will be suppressed). This shows that the Bayesian standard explication of the qualitative notion of confirmation is closely connected to the qualitative notion of correlation.

According to some proponents of confirmation theory, the relation between correlation and confirmation is even stronger. Even the explication of our quantitative notion of confirmation – "evidence E (incrementally) confirms hypothesis H to degree r" – depends on how strongly hypothesis and evidence are correlated, i.e., on the correlation measure presented above. To support this statement this section briefly hints at various confirmation measures that have been suggested in the literature. However, none of these measures are defended. For present purposes it is sufficient to make plausible that confirmation measures and our correlation measure $\mathfrak{corr}$ are closely related. Proposed confirmation measures that link confirmation very intimately to correlation are the following:

Definition 2 (Confirmation 1).

$$r(H, E) = \begin{cases} \log\left[\dfrac{\Pr(H|E)}{\Pr(H)}\right] & \textit{if } \Pr(H|E) > 0 \\ -\infty & \textit{if } \Pr(T|E) = 0 \end{cases}$$

Fitelson (2001) names among others the following advocates of r (or measures ordinally equivalent to it): Horwich (1982), Keynes (1921), and Milne (1996). In this case it is trivial to draw a substantial link between correlation and confirmation.

a coherence measure. In the following no such interpretation is presupposed. Since many philosophers before Shogenji and Wayne have used this or ordinally equivalent measures I refrain from following Wheeler in calling it the Wayne-Shogenji correlation measure. Some of these philosophers are Keynes (1921), Horwich (1982), and Milne (1996). I call this measure the *Simple Correlation* measure since it is considerably simpler than the measure of correlation for finitely many random variables $X_1, \ldots, X_n$ that are usually discussed in the literature on probabilities, such as Watanabe's *Total Correlation* measure C:

$$C(X_1, \ldots, X_n) = \sum_{x_1 \in X_1} \ldots \sum_{x_n \in X_n} \Pr(x_1 \cap \ldots \cap x_n) \times \log\left(\frac{\Pr(x_1 \cap \ldots \cap x_n)}{\Pr(x_1) \times \ldots \times \Pr(x_n)}\right).$$

Corollary 1.

$$r(H,E)=\begin{cases}\log\left[\mathfrak{corr}(H,E)\right] & \textit{if } \mathfrak{corr}(H,E)>0\\ -\infty & \textit{if } \mathfrak{corr}(H,E)=0\end{cases}$$

Another very influential confirmation measure is the following:

Definition 3 (Confirmation 2).

$$l(H,E)=\begin{cases}\log\left[\frac{\Pr(E|H)}{\Pr(E|\neg H)}\right] & \textit{if } \Pr(E|H)>0 \textit{ and } \Pr(E|\neg H)>0\\ \infty & \textit{if } \Pr(E)>0 \textit{ and } \Pr(E|\neg H)=0\\ -\infty & \textit{if } \Pr(E)=0 \textit{ or } \Pr(E|H)=0\end{cases}$$

Among others Fitelson (2001), Good (1960), and Kemeny and Oppenheim (1952) are advocates of l (or measures ordinally equivalent to it). This confirmation measure can also be represented in terms of correlation measure $\mathfrak{corr}$.

Corollary 2.

$$l(H,E)=\begin{cases}\log\left[\dfrac{\mathfrak{corr}(H,E)}{\mathfrak{corr}(\neg H,E)}\right] & \textit{if } \Pr(E|H)>0 \textit{ and } \Pr(E|\neg H)>0\\ \infty & \textit{if } \Pr(E)>0 \textit{ and } \mathfrak{corr}(\neg H,E)=0\\ -\infty & \textit{if } \Pr(E)=0 \textit{ or } \mathfrak{corr}(H,E)=0\end{cases}$$

Even according to the best-known confirmation measure, Carnap's (1950) distance measure of confirmation, confirmation depends on the correlation of the evidence and the hypothesis.

Definition 4 (Confirmation 3).

$$d(H,E)=\Pr(H|E)-\Pr(H)$$

if $\Pr(E)>0$.

Corollary 3.

$$d(H,E)=\left[\mathfrak{corr}(H,E)-1\right]\Pr(H)$$

if $\Pr(E)>0$.

It is to be expected of all confirmation measures that they depend directly on how strongly the hypothesis and the evidence are correlated. This is because confirmation depends on the disparity between either the a priori probability of a hypothesis and its a posteriori probability in the light of the evidence or the a posteriori probability of the hypothesis given the evidence and given the negation of the evidence. From these considerations alone it should be clear that a close study of correlation bears the potential to illuminate our understanding of confirmation.

2.3 Correlation and Coherence

The study of correlation is also of importance to Bayesian coherence theory. According to many Bayesian coherentists, two propositions cohere with each other if and only if they are positively probabilistically correlated (Douven and Meijs 2007; Fitelson 2003; Schupbach 2009; Shogenji 1999). According to some Bayesian coherentists this thought can be generalized. Shogenji (1999) goes furthest by arguing that coherence is nothing else but correlation. Accordingly, Shogenji's (1999) definitions of coherence and of a coherence measure are the following:

Definition 5 (Shogenji Coherence 1).

$$A_1, \ldots, A_n \text{ are coherent if and only if } \mathfrak{corr}(A_1, \ldots, A_n) > 1$$

if $\Pr(A_i) > 0$ for all $i \in \{1, \ldots, n\}$, and 0 otherwise.

Definition 6 (Shogenji Coherence 2).

$$Coh_S(A_1, \ldots, A_n) = \mathfrak{corr}(A_1, \ldots, A_n)$$

if $\Pr(A_i) > 0$ for all $i \in \{1, \ldots, n\}$, and 0 otherwise.

However, some philosophers follow Bovens and Olsson (2000) and Fitelson (2003) in claiming that the simple correlation measure is particularly unsuited to serve as measure of coherence since it is not "sensitive to the (in)dependencies implicit in *all* subsets of" a given set of propositions, and measures of coherence are supposed to be sensitive to these dependencies (Fitelson 2003, 197).[2] In the spirit of this criticism Fitelson (2003) and later Douven and Meijs (2007) suggest defining coherence via confirmation. They hold that a set of propositions is the more coherent the more each subset coheres with all other subsets of the original set. And two sets of propositions cohere with each other more closely, the more the conjunction of the propositions of the first set and the conjunction of the propositions in the second set confirm each

[2] Fitelson expresses the worry most clearly:

> Shogenji's measure is based only on the n-wise (in)dependence of the set E. It is well known that a set E can be j-wise independent, but not i-wise independent, for any $i \neq j$ [...] Shogenji's measure does not take into account the 'mixed' nature of the coherence (incoherence) of E (and its subsets), and it judges E as having the same degree of coherence (incoherence) as a fully independent (or fully dependent) set. (Fitelson 2003, 197)

other. For example (for simplicity, the example is restricted to two propositions) the coherence measure proposed by Fitelson (2003) is defined as follows:

Definition 7 (Fitelson Coherence).

$$Coh_F(A,B) = \frac{1}{2}[\mathcal{F}(A,B) + \mathcal{F}(B,A)]$$

Here $\mathcal{F}$ is Kemeny and Oppenheim's *measure of factual support*, which they introduced in their 1952 joint paper as a measure of evidential support.[3]

Definition 8.

$$\mathcal{F}(A,B) = \frac{p(B|A) - p(B|\neg A)}{p(B|A) + p(B|\neg A)}$$

if $0 < p(A) < 1$ and $p(B) > 0$, otherwise if $p(A) = 0$ or $p(B) = 0$, then $\mathcal{F}(A,B) = -1$ and if $p(A) = 1$ and $p(B) > 0$, then $\mathcal{F}(A,B) = 1$.

Given Fitelson's definition of a coherence measure it is trivial to prove that

Corollary 4.

$$Coh_F(A,B) = \frac{1}{2}\left[\left[\frac{\mathfrak{corr}(A,B) - \mathfrak{corr}(A,\neg B)}{\mathfrak{corr}(A,B) + \mathfrak{corr}(A,\neg B)}\right] + \left[\frac{\mathfrak{corr}(A,B) - \mathfrak{corr}(\neg A,B)}{\mathfrak{corr}(A,B) + \mathfrak{cor}(\neg A,B)}\right]\right]$$

An equally interesting revision of Shogenji's original proposal is presented in Schupbach (2009). According to Schupbach we should follow Fitelson in requiring that the coherence of a set with n elements should not only depend on the n-wise dependence but also on the j-wise dependence of the set's subsets with j elements (for all $j : 1 < j < n$). However, Schupbach also suggests following Shogenji in measuring the j-wise dependence of the set's subsets in terms of the correlation measure $\mathfrak{corr}$, i.e., the Shogenji coherence measure Coh_S.[4]

Furthermore, philosophers like Bovens and Hartmann (2003, 53) and Olsson (2002, 262), who do not define coherence via correlation or confirmation, nevertheless admit that the correlation of two propositions has a positive impact on their coherence. According to them, correlation increases the coherence of the evidence at least *ceteris paribus*. Therefore, understanding correlation is pivotal for any theory of coherence, even if we decide to define coherence via confirmation.

[3] $\mathcal{F}$ is ordinally equivalent to the l measure of confirmation introduced in Sect. 2.2. For a detailed argument in support of l and $\mathcal{F}$ see Fitelson (2001), esp. Sect. 3.2.3.

[4] For a more detailed discussion of Schupbach's measure of coherence and how to render coherence measures sensitive to the correlation of all its subsets see Schupbach (2009).

2.4 Correlation and Explanatory Power

Popper (1959) was one of the first philosophers to suggest a measure for the explanatory power provided by a hypothesis with respect to some evidence. Such a measure should quantify how well a hypothesis explains the evidence. Popper proposes a measure of explanatory power which is ordinally equivalent to the following one by Good (1960)[5]:

Definition 9 (Explanatory Power 1).

$$EP_1(H, E) = \frac{\Pr(E|H)}{\Pr(E)}$$

if $\Pr(H) > 0$ and $\Pr(E) > 0$.

Good (1960) and McGrew (2003) discuss and defend measures of explanatory power that are ordinally equivalent to EP_1. Schupbach and Sprenger (2011) suggest an alternative measure. More specifically, they propose to employ the following measure for measuring the explanatory power provided by a hypothesis regarding some evidence:

Definition 10 (Explanatory Power 2).

$$EP_2(H, E) = \left[\frac{\Pr(H|E) - \Pr(H|\neg E)}{\Pr(H|E) + \Pr(H|\neg E)}\right]$$

if $\Pr(H) > 0$ and $1 > \Pr(E) > 0$.

For the presented measures of explanatory power EP_1 and EP_2, the exact relation to the correlation measure $\mathfrak{corr}$ is given by the following corollaries:

Corollary 5.

$$EP_1(H, E) = \mathfrak{corr}(H, E)$$

$$EP_2(H, E) = \frac{\mathfrak{corr}(E, H) - \mathfrak{corr}(\neg E, H)}{\mathfrak{corr}(E, H) + \mathfrak{corr}(\neg E, H)}$$

if $\Pr(H) > 0$ *and* $1 > \Pr(E) > 0$.

[5] The original formulation of Popper's (1959) measure of explanatory power is this: $EP_P(H, E) = \frac{\Pr(E|H) - \Pr(H)}{\Pr(E|H) + \Pr(H)}$.

3 Correlation and Truth

The preceding section demonstrates the importance of studying correlation. The hope is that the study of correlation will help us to gain a better understanding of various aspects of scientific reasoning which are central to the philosophy of science and epistemology. In particular, suppose that we can explicate all these different aspects of scientific reasoning in terms of correlation or in close relation to it, then the question is whether there is a relation between these forms of reasoning and the primary aim of scientific inquiry: finding the truth. Accordingly, this section focuses on the question whether there is such a connection between truth and correlation.

In fact it is easy to establish a very close link between the truth of a hypothesis and the correlation of that hypothesis and the evidence, by referring to the convergence theorems of, for example, Gaifman and Snir (1982), Schervish and Seidenfeld (1990), or Hawthorne (2011). According to these theorems, the probability of some hypothesis converges to its truth value if the evidence is informative enough to separate the possibilities.[6] Unfortunately, in such a short paper it is not possible to provide a detailed exposition of the mathematically intricate convergence theorems. Accordingly this section will presuppose previous acquaintance with these convergence theorems. I refer the interested reader to the most intelligible exposition of one particular approach to arriving at one of these convergence results, due to Hawthorne (2011). By employing these convergence theorems in the study of the correlation of a given hypothesis and the evidence, the following theorem is provable.

Theorem 1 (Truth-Conduciveness of Correlation). *Let W be a set of possible worlds and let $\mathcal{A}$ be some algebra over W. The elements of $\mathcal{A}$ are interpreted as propositions. Let $e_0, \ldots, e_n, \ldots$ be a sequence of propositions of $\mathcal{A}$ which separates W, and let $e_i^w = e_i$ if $w \vDash e_i$ and $\neg e_i$ otherwise. Let* Pr *be a strict (or regular) probability function on $\mathcal{A}$. Let* $\Pr^*$ *be the unique probability function on the smallest σ-field $\mathcal{A}^*$ containing the field $\mathcal{A}$ satisfying* $\Pr^*(A) = \Pr(A)$ *for all $A \in \mathcal{A}$. Then there is a $W' \subseteq W$ with* $\Pr^*(W') = 1$ *so that the following holds for every $w \in W'$ and all hypotheses H_1 and H_2 of $\mathcal{A}$.*

1. *If $w \vDash H_1$ and $w \vDash \neg H_2$, then:*
   $$\exists n \forall m \geq n : [\mathfrak{corr}(H_1, E_m^w) > \mathfrak{corr}(H_2, E_m^w)]$$
2. *If $w \vDash H_1 \cap H_2$ and $H_1 \vDash H_2$ but $H_2 \nvDash H_1$, then:*
   $$\exists n \forall m \geq n : [\mathfrak{corr}(H_1, E_m^w) > \mathfrak{corr}(H_2, E_m^w)]$$

where $E_m^w = \bigcap_{0 \leq i \leq m} e_i^w$.

This formal result has important ramifications for the application of correlation measures in the philosophy of science and epistemology. In particular, the

[6] A sequence of pieces of evidence separates the set of possibilities W if and only if for every pair of worlds w_i and $w_j \in W$ (with $w_i \neq w_j$) there is one piece of evidence in the sequence such that it is true in one of the possibilities and false in the other.

correlation measure $\mathfrak{corr}$ satisfies the following properties: (i) $\mathfrak{corr}$ favors true hypothesis over false hypotheses and (ii) $\mathfrak{corr}$ favors logically stronger, (i.e., more informative) true hypotheses over logically weaker, (i.e., less informative) true hypotheses after receiving finitely many pieces of evidence and for every additional piece of evidence thereafter.[7]

Theorem 1 implies that if one compares two hypotheses, one of which is true and the other false, then the correlation between the true hypothesis and the evidence is higher than the correlation between the false hypothesis and the evidence (after receiving finitely many pieces of evidence and for every piece of evidence thereafter). Thus, the correlation measure $\mathfrak{corr}$ is truth-conducive in a strong sense: it leads us to true hypotheses after receiving finitely many pieces of evidence. It also shows that if one compares two hypothesis, both of which are true but where one of them is logically stronger, then the correlation between the logically stronger hypothesis and the evidence is higher than the correlation between the logically weaker hypothesis and the evidence (after receiving finitely many pieces of evidence and for every piece of evidence thereafter).[8] This answers the question with respect to the connection between the correlation of a hypothesis and the evidence and the truth of that hypothesis.

However, in many respects the above result leaves us with more questions than answers: What consequences does this result have for the relation between truth, on the one hand, and confirmation, coherence, and explanatory power, on the other? Answering such questions lies at the heart of the Bayesian project in philosophy of science and epistemology. For reasons of space, I cannot here provide an account of how each and every one of the measures of confirmation, coherence, and explanatory power that are discussed in the literature are related to the simple correlation measure $\mathfrak{corr}$. I also cannot discuss in detail which of these measures is the most suitable measure for the given purpose. Nevertheless, I hope to make some remarks that hint at the possible consequences that would ensue if the adequate measures of confirmation, coherence, and explanatory power are indeed among those measures discussed in Sects. 2.2– 2.4

3.1 Correlation, Confirmation, and Truth

According to the three confirmation measures discussed in Sect. 2.2 there is a close connection between confirmation and the correlation measure $\mathfrak{corr}$. Theorem 1 shows that after receiving finitely many pieces of evidence and for every piece of

[7] Note that Theorem 1 restricts these claims: $\mathfrak{corr}$ satisfies both conditions only almost surely: it only holds for every $w \in W'$ where is a $W' \subseteq W$ with $\Pr^*(W') = 1$. It does not necessarily hold for all $w \in W$.

[8] This shows that the correlation measure satisfies two of the three requirements on theory assessment functions put forward in Huber (2008).

evidence thereafter, true hypotheses display a higher degree of correlation with the evidence than false hypotheses. Relying on this result one can prove the following theorem (which can be found in Huber 2008):

Theorem 2 (Truth-Conduciveness of Confirmation). *Let W be a set of possible worlds and let $\mathcal{A}$ be some algebra over W. The elements of $\mathcal{A}$ are interpreted as propositions. Let $e_0, \ldots, e_n, \ldots$ be a sequence of propositions of $\mathcal{A}$ which separates W, and let $e_i^w = e_i$ if $w \models e_i$ and $\neg e_i$ otherwise. Let $\Pr$ be a strict (or regular) probability function on $\mathcal{A}$. Let $\Pr^*$ be the unique probability function on the smallest σ-field $\mathcal{A}^*$ containing the field $\mathcal{A}$ satisfying $\Pr^*(A) = \Pr(A)$ for all $A \in \mathcal{A}$. Then there is a $W' \subseteq W$ with $\Pr^*(W') = 1$ so that the following holds for every $w \in W'$ and all hypotheses $H \in \mathcal{A}$.*

1. *If $w \models H_1$ and $w \models \neg H_2$, then:*
   $\exists n \forall m \geq n : [\mathfrak{c}(H_1, E_m^w) > \mathfrak{c}(H_2, E_m^w)]$, *if* $\mathfrak{c} \in \{r, l, d\}$.
2. *If $w \models H_1 \cap H_2$ and $H_1 \models H_2$ but $H_2 \not\models H_1$, then:*
   $\exists n \forall m \geq n : [\mathfrak{c}(H_1, E_m^w) > \mathfrak{c}(H_2, E_m^w)]$, *if* $\mathfrak{c} \in \{r, d\}$.

where $E_m^w = \bigcap_{0 \leq i \leq m} e_i^w$.

Theorem 2 shows that according to the three confirmation measures discussed in Sect. 2.2, confirmation is truth-conducive: after receiving finitely many pieces of evidence and for every piece of evidence thereafter true hypotheses are confirmed to a higher degree than false hypotheses. In addition we see that confirmation measures r and d distinguish further between true hypotheses as first noted in Huber (2005, 2008). In particular, it shows that if one compares two hypotheses, both of which are true but where one of them is logically stronger, then, after receiving finitely many pieces of evidence and for every piece of evidence thereafter, the logically stronger hypothesis is confirmed to a higher degree by the evidence than the logically weaker hypothesis according to the confirmation measures r and d. The latter result does not hold for the l measure of confirmation which assigns to all true hypotheses the same degree of confirmation in the long run, i.e., the maximum degree of confirmation $+\infty$ (Huber 2005).

For a more detailed discussion of these results see Huber (2005, 2008). For present purposes it suffices to note that we can use Theorem 1 and the intimate connection between correlation and confirmation to learn something about the different confirmation measures. In particular, we can learn whether a high degree of confirmation is an indicator of truth. Moreover, if we adopt a means – end approach to justifying epistemic norms and evaluations these results are highly relevant for the justification of competing confirmation measures. For example, Huber (2005, 2008) argues that confirmation measures such as l that do not favor true informative hypotheses over true but uninformative hypotheses are inadequate as measures of theory assessment, since they do not lead one to the most informative among all true hypotheses.

3.2 Correlation, Coherence, and Truth

Section 2.3 demonstrates that different coherence measures stand in various relations to the correlation measure $\mathfrak{corr}$. Utilizing Theorem 1 we can show that the coherence measures discussed in Sect. 2.3 are truth-conducive as well.

Theorem 3 (Truth-Conduciveness of Coherence). *Let W be a set of possible worlds and let $\mathcal{A}$ be some algebra over W. The elements of $\mathcal{A}$ are interpreted as propositions. Let $e_0, \ldots, e_n, \ldots$ be a sequence of propositions of $\mathcal{A}$ which separates W, and let $e_i^w = e_i$ if $w \vDash e_i$ and $\neg e_i$ otherwise. Let* $\Pr$ *be a strict (or regular) probability function on $\mathcal{A}$. Let* $\Pr^*$ *be the unique probability function on the smallest σ-field $\mathcal{A}^*$ containing the field $\mathcal{A}$ satisfying* $\Pr^*(A) = \Pr(A)$ *for all $A \in \mathcal{A}$. Then there is a $W' \subseteq W$ with* $\Pr^*(W') = 1$ *so that the following holds for every $w \in W'$ and all hypotheses $H \in \mathcal{A}$.*

1. *If $w \vDash H_1$ and $w \vDash \neg H_2$, then:*
   $\exists n \forall m \geq n : [\mathfrak{Coh}(H_1, E_m^w) > \mathfrak{Coh}(H_2, E_m^w)]$, *if* $\mathfrak{Coh} \in \{Coh_S, Coh_F\}$.
2. *If $w \vDash H_1 \cap H_2$ and $H_1 \vDash H_2$ but $H_2 \nvDash H_1$, then:*
   $\exists n \forall m \geq n : [\mathfrak{Coh}(H_1, E_m^w) > \mathfrak{Coh}(H_2, E_m^w)]$, *if* $\mathfrak{Coh} \in \{Coh_S, Coh_F\}$.

where $E_m^w = \bigcap_{0 \leq i \leq m} e_i^w$.

This theorem demonstrates that after receiving finitely many pieces of evidence and for every piece of evidence thereafter true hypotheses cohere to a higher degree with the evidence than false hypotheses.[9] In addition, the coherence measures Coh_S and Coh_F distinguish further between true hypotheses. In particular, if one compares two hypotheses, both of which are true but where one of them is logically stronger, then after receiving finitely many pieces of evidence and for every piece of evidence thereafter the logically stronger hypothesis coheres more with the evidence than the logically weaker hypothesis. Thus the coherence measures of Shogenji and Fitelson can also be considered to be useful tools for judging the acceptability of hypotheses.

However, Theorem 3 does not demonstrate that the measures of coherence proposed by Shogenji and Fitelson adequate measures of coherence. As already said, one recurring objection to Shogenji's proposal is that his coherence measure is not "sensitive to the (in)dependencies implicit in *all* subsets of" a given set of propositions (Fitelson 2003). One prominent objection against Fitelson's measure of coherence is that "if we are confronted with a pair of statements which cannot be false together, Fitelson's function assigns it a coherence value of at most 0. [...] But the fact that one of the assumptions in question must be true does certainly not rule out that they fit together [coherently]" Siebel (2004: 190). Nevertheless the correlation measure $\mathfrak{corr}$ and Theorem 3 are important for the study of coherence since almost all Bayesian coherentists agree that, ceteris paribus, the correlation

[9] Brössel (2008) shows that a similar result can be achieved for the coherence measure suggested by Olsson (2002).

between the propositions of some set is relevant for its degree of coherence. The exact nature of the relation between coherence and correlation, however, can be determined only after we have given an adequate explication of coherence.

3.3 Correlation, Explanatory Power, and Truth

Since Harman (1965), philosophers have been debating whether *inference to the best explanation* is a legitimate form of rational inference. Harman describes this form of inference as follows: “one infers, from the premise that a given hypothesis would provide a ‘better’ explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true” (Harman 1965, 89).

Now let us suppose that one of the proposed measures of explanatory power introduced in Sect. 2.4 indeed gauges the explanatory power provided by a hypothesis with respect to the evidence. Now the question is whether we can justify the inference from the premise that a given hypothesis would provide the best explanation for the evidence (in the sense of these measures), to the conclusion that the given hypothesis is acceptable? This question can indeed be answered positively. With the help of Theorem 1 one can prove that both measures of explanatory power introduced in Sect. 2.4 are truth-conducive in the same way as the measures of confirmation and coherence discussed in the preceding subsections.

Theorem 4 (Truth-Conduciveness of Explanatory Power). *Let W be a set of possible worlds and let $\mathcal{A}$ be some algebra over W. The elements of $\mathcal{A}$ are interpreted as propositions. Let $e_0, \ldots, e_n, \ldots$ be a sequence of propositions of $\mathcal{A}$ which separates W, and let $e_i^w = e_i$ if $w \vDash e_i$ and $\neg e_i$ otherwise. Let* Pr *be a strict (or regular) probability function on $\mathcal{A}$. Let* $\Pr^*$ *be the unique probability function on the smallest σ-field $\mathcal{A}^*$ containing the field $\mathcal{A}$ satisfying* $\Pr^*(A) = \Pr(A)$ *for all $A \in \mathcal{A}$. Then there is a $W' \subseteq W$ with* $\Pr^*(W') = 1$ *so that the following holds for every $w \in W'$ and all hypotheses $H \in \mathcal{A}$.*

1. If $w \vDash H_1$ and $w \vDash \neg H_2$, then:
$$\exists n \forall m \geq n : [\mathfrak{Ep}(H_1, E_m^w) > \mathfrak{Ep}(H_2, E_m^w)], \text{ if } \mathfrak{Ep} \in \{EP_1, EP_2\}.$$
2. If $w \vDash H_1 \cap H_2$ and $H_1 \vDash H_2$ but $H_2 \nvDash H_1$, then:
$$\exists n \forall m \geq n : [\mathfrak{Ep}(H_1, E_m^w) > \mathfrak{Ep}(H_2, E_m^w)], \text{ if } \mathfrak{Ep} \in \{EP_1, EP_2\}.$$

where $E_m^w = \bigcap_{0 \leq i \leq m} e_i^w$.

According to this theorem both proposed measures of explanatory power EP_1, EP_2 take us to true hypotheses. Utilizing the intimate connection between these measures of explanatory power and the correlation measure $\mathfrak{corr}$ one can show that after receiving finitely many pieces of evidence and for every piece of evidence thereafter true hypotheses provide a higher degree of explanatory power than false hypotheses. In addition we see that the measures of explanatory power distinguish further between true hypotheses. In particular, if one compares two hypotheses, both of which are true but where one of them is logically stronger, then after receiving finitely many pieces of evidence and for every piece of evidence thereafter the logically stronger hypothesis displays a higher degree of explanatory power with respect to the evidence than the logically weaker hypothesis. Accordingly it seems that inference to the best explanation is indeed a legitimate form of inference, provided that one of the measures EP_1 and EP_2 is indeed quantifying the degree of explanatory provided by the evidence.

4 Conclusions

Section 2 displays that the simple correlation measure $\mathfrak{corr}$ is indeed the building block of various attempts to capture or explicate essential concepts within philosophy of science and epistemology. In particular, it shows that correlation is closely related to various proposed Bayesian measures of confirmation, coherence, and explanatory power. Section 3 demonstrates how fruitful a detailed investigation of the simple correlation measure $\mathfrak{corr}$ might be if we also relate this investigation to our search for suitable explications for various concepts of scientific reasoning. In particular, Sects. 3.1–3.3 show that the measures of confirmation, coherence, and explanatory power which are discussed most widely in the literature are truth-conducive. They allow us to distinguish between true and false hypotheses after receiving finitely many pieces of evidence and for every piece of evidence thereafter.

However, as already noted, these results come with a caveat. The present paper does not discuss whether the measures of confirmation, coherence, and explanatory power introduced in Sects. 2.2–2.4 are adequate. Accordingly, the exact epistemological consequences that Theorems 1–4 might have depends on the specific theories of confirmation, coherence, etc., that one adopts. If philosophers of science and epistemologists find that confirmation, coherence, explanatory power and other aspects of scientific reasoning are related closely enough to correlation, it might be the case that these important epistemological concepts help us to find the truth. This would be an important milestone for showing that various forms of scientific reasoning can be explicated in terms of probability theory and that we can formulate and justify further epistemic norms and evaluations by relying on these explications.

References

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Bovens, L., & Olsson, E. (2000). Coherentism, reliability and bayesian networks. *Mind, 109*, 685–719.

Brössel, P. (2008). Theory assessment and coherence. *Abstracta, 4*, 57–71.

Carnap, R. (1950). *The logical foundations of probability*. Chicago: University of Chicago Press.

Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese, 156*, 405–425.

Fitelson, B. (2001). *Studies in bayesian confirmation theory*. PhD. Dissertation, University of Wisconsin-Madison (Philosophy).

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis, 63*, 194–199.

Gaifman, H., & Snir, M. (1982). Probabilities over rich languages, testing, and randomness. *Journal of Symbolic Logic, 47*, 495–548.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of The Royal Statistical Society. Series B (Methodological), 22* 319–331.

Harman, G. (1965). The inference to the best explanation. *The Philosophical Review, 74*, 88–95.

Hawthorne, J. (2011). Inductive logic. In Edward Zalta (Ed.), *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/logic-inductive/. Accessed 19 May 2012.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

Huber, F. (2005). What is the point of confirmation? *Philosophy of Science, 72*, 1146–1159.

Huber, F. (2008). Assessing theories, Bayes style. *Synthese, 161*, 89–118.

Kemeny, J., & Oppenheim, P. (1952). Degree of factual support. *Philosophy of Science, 19*, 307–324.

Keynes, J. (1921). *A treatise on probability*. London: Macmillan.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *The British Journal for the Philosophy of Science, 54*, 553–567.

Milne, P. (1996). $log[p(h/eb)/p(h/b)]$ is the one true measure of confirmation. *Philosophy of Science, 63*, 21–26.

Olsson, E. (2002). What is the problem of coherence and truth? *The Journal of Philosophy, 99*, 246–272.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Schervish, M., & Seidenfeld, T. (1990). An approach to consensus and certainty with increasing evidence. *Journal of Statistical Planning and Inference, 25*, 401–414.

Schupbach, J., (2009). New hope for Shogenji's coherence measure. *The British Journal for the Philosophy of Science, 62*, 125–142.

Schupbach, J., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science, 78*, 105–127.

Siebel, M. (2004). On Fitelson's measure of coherence. *Analysis, 64*, 189–190.

Shogenji, T. (1999). Is coherence truth conducive? *Analysis, 59*, 338–345.

Wayne, A. (1995). Bayesianism and diverse evidence. *Philosophy of Science, 62*, 111–121.

Wheeler, G. (2009). Focused correlation and confirmation. *The British Journal for the Philosophy of Science, 60*, 79–100.

The Limits of Probabilism

Wolfgang Pietsch

Abstract I argue that Bayesian probabilism is applicable only to phenomenological theories, in which empirical hypotheses can be clearly distinguished from conventions, while it fails for abstract theories as in physics, where a separation of empirical and conventional parts is usually not feasible. The argument starts from the observation that scientific theories generally contain conventions and that conventions by their very nature cannot be evaluated in terms of probabilities. I then discuss several options how probabilities might be ascribed to a conjunction of empirical hypotheses and conventions – with the result that none of them works. The most promising attempt, namely in terms of probabilities of the empirical consequences given certain conventions, fails due to the mentioned fact that empirical and conventional elements cannot be separated in abstract theories. Thus, Bayesianism cannot provide a foundation for the methodology of abstract sciences.

1 Introduction

In the past decades, there have been various attempts to explicate central concepts in the philosophy of science using methods from probability theory, in particular Bayes' Theorem. I will show in the following that such a Bayesian approach to philosophy of science cannot live up to its task. The simple reason is that it involves a category mistake to ascribe probabilities to theories in the abstract sciences, e.g. in physics. In a nutshell, the argument proceeds as follows. There are three main premises: (1) Besides empirical hypotheses, abstract scientific theories generally contain conventions (Sect. 2.1). (2) To ascribe probabilities to

W. Pietsch (✉)
MCTS, TU München, Arcisstr. 21, 80333 München, Germany
e-mail: pietsch@cvl-a.tum.de

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_5,

conventions constitutes a category mistake (Sect. 2.2). (3) In the abstract sciences, it is impossible to clearly distinguish between conventions and empirical hypotheses (Sect. 2.3). From these three assumptions it follows that one cannot ascribe probabilities to abstract scientific theories as well as to hypotheses, which are either outright conventional or which have an uncertain conventional-empirical status (Sect. 2.4). The result adequately mirrors common practice in the sciences. Bayesian probabilism has proven successful mainly in phenomenological sciences like medicine, psychology, or artificial intelligence while it has only rare applications in physics, chemistry, or other abstract sciences.

2 The Argument from Conventions

Let me first clarify the target of the argument. There exists a wealth of fairly recent literature in the philosophy of science that considers Bayes' Theorem as the ultimate foundation of many aspects of scientific reasoning, including confirmation, belief change, theory reduction, underdetermination, explanation etc. (e.g. Salmon 1966; Rosenkrantz 1977; Horwich 1982; Jeffrey 1983; Earman 1992; Howson and Urbach 2006) In this literature, probabilities are habitually ascribed to scientific theories as the following two examples show. The first concerns the problem of old evidence as discussed in the context of a Bayesian explication of confirmation, for example in Earman and Salmon 1992: "When Einstein proposed his general theory of relativity (H) at the close of 1915 the anomalous advance of the perihelion of Mercury (E) was old news, that is, $Pr(E|K) = 1$ [where K refers to the background knowledge]. Thus, $Pr(H|E.K) = Pr(H|K)$, and so on the incremental conception of confirmation, Mercury's perihelion does not confirm Einstein's theory, a result that flies in the face of the fact that the resolution of the perihelion problem was widely regarded as one of the major triumphs of general relativity." (98) The second example is Jon Dorling's influential Bayesian treatment of the Duhem-Quine problem (1982). Dorling reconstructs episodes mainly from the history of physics in Bayesian terms involving probabilities for physical theories like orthodox quantum mechanics, local hidden variable theories, Newtonian gravity, or general relativity. I will now proceed to show that such discussions cannot provide much methodological insight since they are based on the mistaken assumption that physical theories can be ascribed probabilities.

2.1 *Abstract Scientific Theories Contain Conventions*

While the notion of convention has received its fair share of philosophical attention, it remains a notoriously difficult concept. Fortunately, we do not have to delve into the details of the debate since the two features of conventions that will be relevant for the argument against probabilism are largely uncontroversial: to each choice of a convention there exists (i) at least one viable alternative which is (ii) incompatible with the other choices. To substantiate this claim let us briefly look at two widely influential accounts by David Lewis and Henri Poincaré.

Arguably, the classic philosophical treatment of convention in the twentieth century is due to David Lewis (2002). Lewis gives a somewhat lengthy definition in terms of a regularity in the behavior of members of a community resulting from a coordination problem within this community (78). Essentially, if most members conform to a certain regularity, then the preferences of all members should be such that everyone conforms to this regularity. The behavior of the majority thus determines the preferences of the individual member. According to Lewis, if the majority had opted for a different regularity, the individual members would have had to follow along. Lewis stresses that there is "no such thing as the only possible convention" and that "it is redundant to speak of an arbitrary convention" (70). Thus, Lewis' definition implies (i) and (ii), since there are always alternative choices which are mutually incompatible, because the different kinds of behavior are incompatible. Typical everyday conventions like dress codes or traffic rules fit well with Lewis' definition but also scientific conventions like the choice of base units.

In comparison with Lewis' account, Henri Poincaré has a much narrower focus on the role of conventions in science and especially in physics:

> Are [conventions in physics] arbitrary? No; for if they were, they would not be fertile. Experience leaves us our freedom of choice, but it guides us by helping us to discern the most convenient path to follow. Our laws [when of conventional nature] are therefore like those of an absolute monarch, who is wise and consults his council of state. Some people have been struck by this characteristic of free convention which may be recognised in certain fundamental principles of the sciences. Some have set no limits to their generalisations, and at the same time they have forgotten that there is a difference between liberty and the purely arbitrary. (Poincaré 1905, p. xxiii)

While Poincaré also stresses the existence of incompatible alternatives, he is less clear in comparison with Lewis to which extent the choice between alternatives depends only on the behavior of the majority. For example, Poincaré pointed out the conventional nature of the axioms of geometry, but notoriously claimed that there is one best choice, namely Euclidean geometry, which Poincaré singles out in terms of 'commodité', i.e. convenience. While Poincaré agrees with Lewis that there is always some freedom, the specific choice of a convention need not be arbitrary. Still, Poincaré's account implies that there is at least one viable, incompatible alternative to each convention which might however fare worse in terms of convenience. Thus, the characteristics (i) and (ii) are fulfilled, which is enough for the argument in the next Sect. 2.2 to go through.

Let me illustrate these properties by means of the convention that length is measured in meters. Clearly, there is a variety of other units of length that would be equally good, for example measuring length in feet. There even exists an infinite number of alternative choices, if meter is multiplied with an arbitrary number and the result taken as the base unit. These choices are incompatible since we have to decide, which numbers to write on measuring rods, maps and traffic signs and to use for calculations of length in scientific or everyday contexts. Thus, in agreement with Lewis' account, the members of the relevant community have to conventionally fix the unit of length in the interest of a common goal, namely communicating distances. Then, they must act according to the established convention, resulting in regularities of behavior. Obviously, there are important contextual and pragmatic factors that narrow down sensible choices of the unit of length. It is no coincidence that the fundamental unit is of the order of the human body size. After all, most problems that we deal with in every-day life are of that order of magnitude. Finally, one certainly cannot speak of meter being the right or wrong or even a probable base unit. Rather, it is a suitable or convenient choice with respect to certain applications.

There is no shortage of conventions in scientific theories. Let me list a number of different types without being exhaustive: (1) Whenever a measurable continuous quantity forms a part of a scientific theory, a convention is necessary to determine the unit of this quantity. That we measure length in meters is an example, or energy in joules, time in seconds etc. (2) Conventions are also required when scientific theories with invariances or symmetries are applied. Take for example homogeneity in space as a symmetry (or invariance) of physical theories. Whenever a specific set-up is considered, the symmetry is broken and a suitable origin of the spatial coordinate system must be introduced. The choice of the origin is of course conventional. Another example concerns the Lorentz-invariance of the theory of special relativity. For specific calculations, a convention must be introduced that fixes the velocity of the observer and thereby the choice of inertial system. (3) Often, concepts are introduced in scientific theories in terms of suitable conventional definitions. As an example, Newton's second axiom has frequently been interpreted as a definition of the (very useful) quantity of force from mass and acceleration. On the basis of this example, let me briefly provide some plausible considerations that in fundamental theories the conventional part includes several of the core statements of the theory. The reason is that fundamental theories themselves introduce the language required to describe the respective phenomena. They thus serve a linguistic function and must generally contain explicit or at least implicit definitions of the core terms of the theory. Now, these definitions are at least in certain aspects of conventional nature. Since they often link central concepts of the theory, they plausibly belong to the core of fundamental theories, as in the case of Newton's second axiom.

Items (1) and (2) show that conventions have to be introduced whenever a theory is applied to a specific physical set-up. Thus, a theory can only be confirmed or disconfirmed by observations if conventions are taken into account. Confirmation thus never concerns the bare structure of the theory itself, but always the structure plus a suitable choice of conventions. Item (3) suggests that fundamental theories will contain conventional elements that concern the very core of these theories, because they themselves introduce the vocabulary necessary for an adequate description of the respective phenomena.

2.2 No Probabilities for Conventions

I will now argue against ascribing probabilities to conventions. The argument, which holds independently of a specific interpretation of probability, is based on the concept of convention as spelled out in the previous section, in particular on the two crucial characteristics that were pointed out: namely that to every convention there is (i) at least one alternative choice, which is (ii) incompatible with the other choices.

Consider a certain choice C_1 of a convention, e.g. $C_1 =$ 'the unit of length is one meter'. Two cases are distinguished, both of which run into contradictions. In the first case, the probability of the convention is assumed to be considerably smaller than one, i.e. $P(C_1) = p < < 1$. The problem with this option is that no theory or more generally set of propositions T containing the convention C_1 can attain a probability that is larger than p since: $P(T) = P(T\backslash C_1|C_1)\ P(C_1) \leq p$, where $T\backslash C_1$ refers to the set of propositions T without the convention C_1. In other words, there would be no well-confirmed theories if the probabilities of conventions would be considerably smaller than one.

In the second case, if the probability is approximately 1, i.e. $P(C_1) = p \approx 1$, then all other choices C_2, C_3, etc. must have a probability close to zero. This can be derived easily from the axioms of probability: $P(C_1 \text{ v } C_2) = P(C_1) + P(C_2) - P(C_1\ \&\ C_2) = P(C_1) + P(C_2) \leq 1$, from which follows that $P(C_2) \leq 1 - p \approx 0$. In the derivation, the fact was used that C_1 and C_2 are incompatible choices and therefore $P(C_1\ \&\ C_2) = 0$. By this reasoning the probabilities of all alternative choices C_2, C_3, etc. must be close to zero, and therefore any theory or set of statements containing these conventions would have a probability close to zero. But this contradicts the second assumption that at least one alternative C_2 is a viable choice. For example, one can very well construct a highly plausible theory containing the convention that "the unit of length is one foot".

The problems remain even if a specific choice of convention is considered to be the best or in Poincaré's words the most convenient choice. Essentially, there is no reason to suppose that a measure of convenience will obey the probability axioms. For example, one cannot reasonably assume that the degree of convenience of all possible alternatives should add up to one. After all, if the degree of convenience of a particular choice increases due to some new information or a change in context, why should the degree of convenience of the other choices decrease accordingly? One may well imagine that some new information leads to an increase or decrease in the convenience of *all* possible alternatives.

Also, it is often the case that a convention has an infinite number of alternatives. If the measure of convenience is normalized to one, as required by the axioms of probability, and a reasonably smooth curve for the convenience measure is assumed, i.e. that similar choices of conventions should not differ significantly in terms of convenience, then we must ascribe to all choices of convention a degree of convenience of zero. This consequence is absurd, since intuitively at least some plausible choices should have a finite convenience different from zero. For example, the convenience of measuring length in meter is certainly finite, even though there exists an infinite number of alternatives of approximately similar convenience. This problem is specific to conventions and does not occur for empirical statements due to a crucial difference: the values of conventional quantities are always exact, while the values of empirical quantities come equipped with an error function. Consider for example the proposition P: ‘quantity X has a value of 1.3’. If P is of conventional nature, then *all* further digits of X are fixed to be zero by convention. Therefore, the measure of convenience of P will always be zero, as long as the convenience function over possible alternative values for X remains smooth and continuous. By contrast, if P is of empirical nature, an error function is automatically associated with X. In other words, the proposition P then implies that X can take on different values within a certain interval, say between 1.250 and 1.349. Taking into account this range of error, the probability of P can be finite and even approach one, although the distribution over alternative values remains smooth and continuous. Thus, a problem arises with the updating process for conventions. In the case of smooth and continuous distribution functions, the convenience of certain choices of conventions will always stay zero, while the probability of empirical statements can take on any value between zero and one. In addition, there may be some technical difficulties regarding a supposed Bayesian updating of continuous convenience functions, since Bayes’ Theorem usually works with discrete distributions. However, these difficulties could presumably be overcome by employing more sophisticated techniques. In summary, any convenience measure that is supposed to handle cases with an infinite number of alternatives cannot be normalized to one and therefore will not satisfy the probability axioms. Not surprisingly, convenience is not probability.

As a last resort, maybe probabilities should not be ascribed to particular choices of conventions but to equivalence classes of conventions. But it is generally impossible to choose equivalence classes such that the probability axioms are satisfied. Consider once more the measure of length. In order to solve the problems mentioned above, the equivalence class should comprise at least all base units that differ from meter by a simple factor. Also, this equivalence class should be given probability one. However, one can easily imagine more complex choices of measure where the base unit relates to the meter in terms of a space- and time-dependent function (resulting essentially in non-Euclidean geometries). Should these complex choices all be given a probability zero? Are they always and in all contexts less convenient then the ordinary meter and its equivalence class? Into which equivalence classes should we partition these further measures of space? In the end, these difficulties, which mainly result from the impossibility to draw a clear line between the conventional and the empirical in abstract theories, are insuperable. Equivalence classes cannot provide a basis for ascribing probabilities to conventions.

2.3 Criticizing the Conventional-Empirical Distinction

In the phenomenological sciences that deal with more or less directly observable events it is not difficult to formulate ‘theories’ that are purely empirical. However, in abstract and fundamental sciences like physics this is not possible since as we had seen in Sect. 2.1, conventions usually concern some of the core elements of the theory. Also, a clear distinction between the empirical and the conventional part of such theories cannot be drawn – resulting in considerable disagreement between different scientists about the empirical or conventional nature of certain propositions. Often, there is no disagreement about the propositions themselves or about their formulation but only about their empirical or conventional status. Furthermore, the *same* scientist may consider a proposition empirical in some contexts and conventional in others. This point is of course closely related to well-known criticism of the analytic-synthetic distinction.[1]

Let me present some examples. (1) The choice of measure for fundamental quantities is often much more complex than suggested in the discussion concerning the unit of length in Sect. 2.1. This can already be deduced from the fact that large amounts of money continue to be invested in metrology, i.e. the science of measuring. Fundamental units continue to be redefined – always connected with a shift in the empirical or conventional status of fundamental propositions.[2] Generally, what is at stake in choosing a fundamental unit is more than just a simple factor as in: 1 m = 3.28 ft. Rather, the choice of measure is much more complex, and is deeply interwoven with the progress in the respective science itself. Chang (2007) demonstrates this in a very detailed historical case study using the example of temperature.

Another good example concerns the debate regarding the conventionality of geometry beginning in the second half of the nineteenth century. Around that time, mathematicians realized that non-Euclidean geometries, i.e. geometries which do not obey the parallel postulate, can be formulated in a consistent way. Soon mathematical physicists like Hermann von Helmholtz and Henri Poincaré realized that such non-Euclidean geometries can be used for representing physical space and the motion of particles therein if complementary changes are introduced for the physical laws (e.g. von Helmholtz 1870). This insight resulted in the thesis of the conventionality of geometry, i.e. that there is a number of possible choices (Euclidean and various non-Euclidean) how the geometry of physical space can be consistently formulated – each involving a different choice of measure and corresponding changes in the fundamental laws of physics. Even though Einstein still in Einstein (1921) held that the thesis of conventionality of geometry was in

[1] Essentially, I agree with Duhem that difficulties with the analytic-synthetic distinction are relevant mainly for abstract sciences, not so much for phenomenological sciences. This underdetermination in empirical-conventional content of abstract theories serves an important function for scientific progress, as will be shown on another occasion.

[2] An interesting recent development in metrology aims at the redefinition of four of the seven international base units, namely the kilogram, the ampere, the kelvin, and the mole. The new definitions will rely on fixing four fundamental constants, namely the Planck constant, the elementary charge, the Boltzmann constant, and the Avogadro constant, respectively. Strictly speaking, these constants will be turned into conventions. For a philosophical analysis, see Pietsch 2013.

principle correct, there has been a tendency to empiricize the geometry of physical space after the advent of general relativity. Over the last hundred years the debate concerning the empirical or conventional nature of geometry has continued without a clear result, which is hardly surprising given the complexity of the issue.

(2) A related issue concerns the empirical or conventional nature of constants as can be illustrated using the vacuum speed of light. The question if this constant is empirical or conventional depends on the stance that one assumes towards the relation between space and time. One might, à la Minkowski, insist that we ultimately live in a four-dimensional space-time, i.e. that space and time are just different dimensions of one and the same entity. From this viewpoint, it is a historical coincidence, resulting from a premature understanding of physics, that we happen to measure space and time with different units. In principle, leaving aside pragmatic considerations, one should use the same measure for time as for space. Consequently, the velocity of light is a mere convention.

To ask if a conventional constant can change over time is nonsensical. Nevertheless, physicists take seriously such a possibility on cosmological scales, obviously denying a strong conceptual identity of space and time. Still worse, the same physicists sometimes treat the velocity of light as a convention in certain contexts, for example when dealing with events in terrestrial laboratories, but might be willing to concede some limited empirical content when considering astronomical scales. Thus, there is no general agreement if the velocity of light is an empirical or conventional constant and given the complex ramifications with immensely difficult questions like the conceptual nature of space-time, it is not very plausible that there will ever be a definite answer.

(3) The empirical-conventional distinction is also blurred when it comes to the question which quantities are fundamental in a theory and which are secondary or merely defined. Consider again Newton's second axiom/law as an example: force $=$ mass $\times$ acceleration. Throughout the history of physics, the exact status of this axiom has been debated. The conventional-empirical status of the second axiom obviously depends on the intricate issue which of the quantities figuring in the second axiom are fundamental and which are not.

The fixing of measure for fundamental quantities, the determination of fundamental constants, or the determination of which quantities are fundamental and which derived – all these issues are tasks rather for abstract sciences than for phenomenological sciences since the latter mostly rely on a language determined by other (abstract) sciences.

2.4 Bringing Together the Argument

On the basis of the premises 2.1–3, I will now argue that abstract theories and hypotheses cannot be ascribed probabilities. Emphatically, a conception of theories as a conjunction of empirical hypotheses, as is prevalent in some of the Bayesian literature, is too simplistic for the abstract sciences. Rather, as shown in

Sect. 2.1, abstract scientific theories generally involve conventions besides empirical hypotheses. They may of course contain still other elements which might also not be probability bearers but here it suffices to focus on conventions.

Consider a toy theory t made up of hypotheses $h_1, \ldots, h_n$ and conventions $c_1, \ldots, c_m$. There are several ways how one could ascribe probabilities to such a theory: (i) $P(t) = P(h \& c) = P(h|c) P(c) = P(c|h) P(h)$; (ii) $P(h|c)$, which allows for an inverse probability $P(c|h)$; (iii) $P_c(h)$, which does not allow for an inverse probability.

The first option (i) must be excluded since it involves ascribing probabilities to conventions either in P(c) or in P(c|h) while in Sect. 2.2 we have shown that this constitutes a category mistake. Option (ii), the probability of the empirical hypotheses given certain conventions, must be excluded for the same reason that probabilities are ascribed to conventions. After all, according to the definition of conditional probability, we have $P(h|c) = P(h \& c)/P(c)$.

Thus, the premises $2.1 + 2.2$ imply that probabilities cannot be ascribed to the whole set of propositions of an abstract theory nor to those statements or hypotheses in the abstract sciences which are either outright conventional or have an uncertain conventional-empirical status. Crucially, as we saw in Sect. 2.1, the conventional part generally comprises core elements of abstract theories. Thus, a Bayesian approach to philosophy of science must at least be reoriented or refined with respect to what is meant by the probability of abstract theories or hypotheses.

The plausible candidate for making sense of such probabilities is option (iii), i.e. $P_c(h)$, which also refers to the probability of the empirical hypotheses given certain conventions, but unlike in option (ii), inverse probabilities $P_h(c)$ are not allowed. Thus, probabilities for conventions are avoided. Apparently, this option works well in the phenomenological sciences like medicine, psychology etc., where hypotheses are often purely empirical and ‘theories’ just conjunctions of empirical hypotheses. In the most benign cases, the probabilities of such purely empirical theories or hypotheses are independent of the specific choice of conventions, i.e. $P_c(h) =: P(h)$. In more malicious cases, probabilities change with different choices of conventions.[3]

However, option (iii) fails if a distinction between empirical hypotheses and conventions cannot be drawn. As we have laid out in Sect. 2.3, this is the case in abstract sciences like physics. Essentially, physical theories have functions beyond making assertions about the world, they also provide an adequate language for speaking about physical phenomena by introducing the necessary vocabulary in terms of conventional definitions. It is with respect to the first function that the notion of probability makes sense but not with respect to the second function. Since both functions are inextricably intertwined in physics, one cannot speak of the probability of a physical theory. Also, single propositions in the abstract sciences

[3] Confer discussions of language change for example in Williamson (2003) or Romeyn (2005, Chap. 8.6).

often serve both an empirical and a definitional-conventional purpose and therefore cannot be ascribed probabilities either. Arguably, this holds for many axioms in physics like the Newtonian axioms or the axioms of relativity theory including the constancy of the speed of light, as was discussed in Sect. 2.3.

A Bayesian probabilist may nevertheless insist to identify the probability of an abstract theory with the probability of its empirical consequences $P_x(h)$, where x contains both the clearly conventional part of the theory and those propositions that have a doubtful status. Note again that the empirical content of a theory can shift with different choices of conventions. A concrete example is given in Pietsch (2013) concerning different interpretations of the Fizeau-Foucault experiment to measure the speed of light (Sect. 3.2). Furthermore, probability as a measure of confirmation will concern only the empirical consequences of the theory, never the entire theory including x. For example, Bayes' Theorem would read: $P_x(h|e) = P_x(e|h) P_x(h)/P_x(e)$. Since x functions merely as an index, reconstructions of methodological concepts relying on this version of Bayes' Theorem could never provide much insight regarding the crucial role of x.

These difficulties could possibly be avoided if one assumed that the observational consequences uniquely implied the conventional part x. However, in scientific practice, this never seems to be the case. Also, many interesting questions concerning the relation between the empirical and the conventional would automatically be suppressed. Indeed, many fundamental concepts in the philosophy of science concern exactly the definitional-conventional function of abstract scientific theories in relation to the empirical basis. For example, underdetermination is about different descriptions of the same phenomena which stand in a non-trivial relation with each other. Holism is partly about different perspectives on what terms are fundamental and what terms are defined. Theory reduction is about connecting different languages, usually macro and micro, and so on. All these methodological concepts thus cannot be explicated in probabilistic terms.

3 Conclusion

Degrees of belief in abstract theories or abstract hypotheses cannot be spelled out in terms of probabilities, not even in terms of qualitative probabilities. In a sense, 'belief' in abstract theories has a passive and an active component: A passive, evidential component referring to empirical facts and an active, conventional component denoting the willingness of a scientist to stick to certain propositions. The first can be spelled out in terms of probabilities, the second cannot. If these passive and active components cannot sensibly be separated, as is the case for abstract theories or hypotheses, then probabilities cannot be ascribed and a Bayesian approach is not feasible.

References

Chang, H. (2007). *Inventing temperature. Measurement and scientific progress*. Oxford: Oxford University Press.

Dorling, J. (1982). *Further illustrations of the Bayesian solution of Duhem's problem*. http://www.princeton.edu/~bayesway/Dorling/dorling.html. Accessed 18 July 2012.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

Earman, J., & Salmon, W. (1992). The confirmation of scientific hypotheses. In M. H. Salmon et al. (Eds.), *Introduction to the philosophy of science* (pp. 42–103). Englewood Cliffs: Prentice Hall.

Einstein, A. (1921). Geometry and experience. In *Sidelights on relativity* (pp. 25–56). Mineola: Dover (2010).

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning. The Bayesian approach* (3rd ed.). Chicago: Open Court.

Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Chicago: University of Chicago Press.

Lewis, D. (2002). *Convention*. Oxford: Wiley-Blackwell.

Poincaré, H. (1905). *Science and hypothesis*. London: Walter Scott.

Pietsch, W. (2013). A revolution without tooth and claw – Redefining the physical base units. http://www.wolfgangpietsch.de/pietsch-new_si.pdf. Accessed 28 Sept. 2013.

Romeyn, J.-W. (2005). *Bayesian inductive logic*. Alblasserdam: Haveka BV.

Rosenkrantz, R. (1977). *Inference, method, and decision: Towards a Bayesian philosophy of science*. Dordrecht: Reidel.

Salmon, W. (1966). *The foundations of scientific inference*. Pittsburgh: University of Pittsburgh Press.

von Helmholtz, H. (1870). Über den Ursprung und die Bedeutung der geometrischen Axiome. In *Populäre Wissenschaftliche Vorträge von H. Helmholtz. Drittes Heft* (pp. 21–54). Braunschweig: Vieweg und Sohn (1876).

Williamson, J. (2003). Bayesianism and language change. *Journal of Logic, Language and Information, 12*(1), 53–97.

Part II
Philosophy of Science: Idealization, Representation and Explanation

How Organization Explains

Jaakko Kuorikoski and Petri Ylikoski

Abstract Constitutive mechanistic explanations explain a property of a whole with the properties of its parts and their organization. Carl Craver's mutual manipulability criterion for constitutive relevance only captures the explanatory relevance of causal properties of parts and leaves the organization side of mechanistic explanation unaccounted for. We use the contrastive counterfactual theory of explanation and an account of the dimensions of organization to build a typology of organizational dependence. We analyse organizational explanations in terms of such dependencies and emphasize the importance of modular organizational motifs. We apply this framework to two cases from social science and systems biology, both fields in which organization plays a crucial explanatory role: agent-based simulations of residential segregation and the recent work on network motifs in transcription regulation networks.

1 Introduction

Mechanistic explanation has been identified as an important type of scientific explanation (Glennan 2002; Craver 2007; Hedström and Ylikoski 2010). Constitutive and developmental mechanistic explanations explain a property of a system with the properties of its parts and their organized interaction. The representation of spatial environment in rats is explained by the activities and mutual organization

J. Kuorikoski (✉)
Social and Moral Philosophy/Department of Political and Economic Studies, University of Helsinki, P.O. Box 24, 00014 Helsinki, Finland
e-mail: jaakko.kuorikoski@helsinki.fi

P. Ylikoski
Department of Social Research, University of Helsinki, P.O. Box 54, 00014 Helsinki, Finland
e-mail: petri.ylikoski@helsinki.fi

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_6,

of neurons in the hippocampus and the efficient allocation of goods in a market system is explained by the properties and the structure of interaction of the agents participating in the market. Thus far, everyone agrees. It is also agreed that the organization of the parts has a crucial role in these explanations. However, there is very little discussion of organization as an explanatory variable. Most accounts of mechanistic explanation simply treat it as a stable background condition. Thus, the challenge for the analysis of mechanistic explanation is to move beyond the mere acknowledgement that the organization of the parts is important for the behaviour of the whole.

The lack of analytical tools with which to approach the explanatory import of organization is not a trivial lacuna. The difficulties in conceptualizing the role of organization have in the past often manifested themselves in metaphysical vocabulary, such as claims of emergence or irreducibility. Although the philosophical credibility of such notions has diminished in the wake of the rise of the mechanistic philosophy of science, and their role as simple placeholders for a lack of understanding is now widely acknowledged, simply replacing the word ‘emergence’ with the word ‘organization’ is not enough to fill the gap in the understanding of how the whole can be more than the sum of its parts. Without an account of *how* organization explains, such talk amounts to no more than a transformation of emergence mysticism into ‘organization mysticism’.

This paper is an attempt at filling this gap in the literature on mechanistic explanation. We will begin by briefly presenting an account of explanation that will serve as the basis for our account of organizational explanation. Then we will argue that William Wimsatt’s idea of emergence as a failure of aggregativity, which Carl Craver, among others, takes to be the most promising starting point for an analysis of the role of organization, does not as such provide an appropriate scheme for analysing organization’s explanatory role. In place of Wimsatt’s scheme, we suggest three dimensions of organizational dependence: diversity in the kinds of components, the network structure between the components, and diversity in the kinds of relations. Furthermore, we argue that understanding organization as an explanatory variable proceeds best by first understanding relatively simple organizational motifs. We conclude the paper with two compact illustrations of this idea: agent-based simulations of residential segregation and network motifs in gene regulation.

2 A Toolbox for Understanding Mechanistic Explanation

We take the contrastive-counterfactual theory of explanation (Woodward 2003; Ylikoski and Kuorikoski 2010) as our starting point. Explanation consists of tracking and exhibiting dependencies. Dependencies differ from regularities in that they are to be analysed in terms of counterfactual conditionals and thus have an irreducibly modal component: A explains B only if, if A had been different, B would have been different as well. Explanations thus show what makes or made a difference to the thing to be explained. The *relata* of explanations are values of variables, and explanations are therefore doubly contrastive, the ranges of the possible values of the *explanandum* and the *explanans* variables forming the relevant contrast classes (Woodward 2003).

The criterion of explanatory relevance is counterfactual dependence, but not all counterfactual inferences reflect relations of dependence in the world, rather than epistemic or inferential relations between our representations of the world. The incoming weather front explains the change in the barometer reading since it causes it, but it also makes (a sort of) sense to conclude that if the barometer reading had stayed high, there would not have been a low-pressure system approaching. Yet although the barometer reading is a reason to believe in the presence of low pressure, it does not explain it. We therefore need an additional element in the contrastive-counterfactual framework: that of an intervention. Variables are causally related if we could (in principle) bring about changes in one variable by intervening on the other. Intervention is a causal manipulation of a single variable which is not itself caused by (or even correlated with) anything within the system and breaks or bypasses other causal influences of the target variable but does not directly affect any other variables or dependency relations (Woodward 2003). With the concept of intervention, we can now distinguish causal dependencies from inferential dependencies, define causal order, disambiguate closely related causal concepts and clarify causal reasoning in complex causal structures.

The final element in our toolkit is the concept of understanding. We operationalize understanding as the ability to make correct counterfactual inferences concerning the consequences of interventions on the phenomenon to be understood (Ylikoski 2009; Ylikoski and Kuorikoski 2010). Such inferential ability is based on knowledge of causal and constitutive dependencies, but is not the same thing. Understanding is not simply a matter of possessing knowledge, but of proficiency in using it to make inferences beyond what has actually happened. Understanding should therefore be sharply distinguished from the psychological sense of understanding, which may or may not accompany any increase in understanding (Keil 2003). The kind and degree of understanding created by an explanation, and thus something that might be called its "explanatory power", can therefore be spelled out by listing the *what-if-things-had-been-different* questions (w-questions) answerable with the conveyed explanatory information, taking into account the cognitive limits of the agents engaged in seeking and giving the explanation (Ylikoski 2009).

3 Constitutive Explanation and Manipulation

The interventionist theory of explanation is an account of causal explanation, but the mechanistic explanation of a property of a whole in terms of its parts and their organization is not causal, strictly speaking, since the parts are constitutive of the whole and therefore cannot be independently manipulated. The parts and the whole that they constitute are not independent existences. Also, the determination relation between the properties of the parts and the whole is not a process in time. Despite these differences, as Craver (2007) has suggested, the concept of intervention is also useful for analysing the process by which we investigate the explanatory relevance of particular parts (and their properties) for the properties of the system. Craver proposes that the correct criterion of constitutive explanatory relevance is one of *mutual manipulability*: a property of a component part is explanatorily relevant if and only if by intervening on the system-level property, we induce changes in the property of the component, and by intervening on the component, induce changes on the system level.

Craver's criterion certainly fits well with our epistemic practices in that mutual manipulation has a crucial role in learning about constitutive relations. Lesioning or stimulating parts of the brain and observing the effects that these interventions have on the overall functioning and, conversely, observing localized activation patterns within the brain while the subject is undertaking some cognitive tasks (system-level intervention) both provide evidence of the constitutive relevance of the properties of parts. While Craver is overplaying the symmetry of constitutive relevance, and there might be some problems with his definition of intervention, his account is an important contribution to the theory constitutive explanation. His account is also fully compatible with the contrastive-counterfactual approach to explanation that we are advocating.

The contrastive-counterfactual approach is therefore applicable to the analysis of mechanistic explanations: a causal property of a component part contrastively explains an aspect of a property of the whole mechanism if intervening on it would change the property of the whole from its actual value to a contrast value (or vice versa). Knowledge of such constitutive dependencies provides understanding of why the mechanism behaves as it does by grounding answers to w-questions concerning the effects of possible interventions on the component parts. The problem now is to expand this analysis to the all-important explanatory relevance of the organization of the parts.

The basic idea of extending the contrastive-counterfactual analysis to the role of organization is straightforward enough: organization constitutively explains why a system-level property is p rather than p' iff by appropriately intervening on the organization, the property would change to its contrast value. The problem is that conceiving organization simply as an additional explanatory variable is not very illuminating and does not really move us beyond emergence mysticism. After all, the additional explanatory variable could have just as well been labelled 'emergence': if there had not been emergence, the property of the whole would have been different as well.

4 Organizational Dependence as Non-Aggregativity

The main problem in analysing organizational explanation in terms of an organization variable is that there is no single unique form of organizational constitutive dependence. What we would like to have is a taxonomy of kinds of ways in which the property of the whole depends on the organization of the component parts. Craver also acknowledges that additional conceptual tools are needed to understand the explanatory relevance of organization and suggests that these could be derived from William Wimsatt's (2007) account of emergence as non-aggregativity (Craver 2007, p. 135). Wimsatt explores conditions under which the whole is literally nothing more than the sum of its parts. Failures of such conditions therefore mark ways in which the whole is dependent on the organization of its parts. Wimsatt's four conditions for aggregativity are the following (2007, pp. 280–281):

1. **IS** (***Inter* S***ubstitution*): Invariance of the system property under operations rearranging the parts in the system or interchanging any number of parts with a corresponding number of parts from a relevant equivalence class of parts (cf. the commutativity of the composition function).
2. **QS** (*Size scaling*): The **Q**ualitative **S**imilarity of the system property (identity, or if a quantitative property, differing only in value) under the addition or subtraction of parts (cf. the recursive generability of a class of composition functions).
3. **RA** (Decomposition and ***ReA****ggregation*): Invariance of the system property under operations involving the decomposition and reaggregation of parts (cf. the associativity of the composition function).
4. **CI** (*Linearity*): There are no **C**ooperative or **I**nhibitory interactions among the parts of the system which affect this property.

The idea would then be that whenever there is a failure of one of these conditions, there is a specific type of dependency between the organization and the property of the whole and that this dependence is invariant under interventions and thus grounds answers to w–questions. Wimsatt's conditions could therefore serve as a basis for analysing the constitutive explanatory relevance of organization.

There are two problems with this suggestion. First, trying to translate Wimsatt's conditions into invariant dependencies reveals that the conditions conflate properties of representations and properties of the represented system. Thus, a direct translation leads to dependencies that are a mix of ontic and inferential dependencies. As an example, let us look at the first condition, IS. The formal idea expressed with the concept of the composition function is clear: changing the order of the arguments of the function does not affect the result. The idea also works well in Wimsatt's favourite example of amplifiers: the order of serially connected amplifiers does not affect (approximately) the total amplification ratio (2007, p. 285). In this case the properties of the representation and the thing represented go nicely hand in hand: changing the arguments is a formal operation corresponding to a physical intervention of changing the order of amplifiers. But this is a special case. If the organization is even slightly more complex (i.e., not serial), the correspondence between commutativity and the irrelevance of the way in which the components are 'ordered' breaks down. The same basic problem also haunts the condition RA: there is a sense in which decomposition and re-aggregation can be thought of as physical causal operations, but this sense is clearly different from Wimsatt's formal notions expressed in terms of the composition function. If we cannot keep physical and conceptual 'interventions' separate, it follows that we cannot distinguish between conceptual exploration based on inferential dependencies between representations and genuine explanation based on ontic dependencies between things in the world.

The second problem with Wimsatt's conditions is that the failures are more akin to symptoms of the role of organization rather than an analysis of the thing itself. The conditions list different types of cases in which we cannot simply aggregate the whole from its parts, but they do not really explicate *why* this cannot be done. What we need is a more general and analytically fruitful way of conceiving organization as an explanatory variable. We next propose three such dimensions of organizational dependence and introduce the crucial concept of organizational motif, which links differences along these dimensions to differences in the property of the whole.

5 Dimensions of Organizational Dependence

We approach the taxonomy of organizational dependence by first asking what organization is made of. By finding the basic constituents of organization, we find the things that, if changed, would lead to changes in the property of the whole. We follow Wimsatt in starting from the limiting case of the complete lack of organization and then build up from there. First, if a system is to have any internal organization, it has to possess some internal differentiation: if the whole is to have parts, it has to have some features according to which the boundaries of those parts can be delineated. In the simplest aggregative case, the parts are all alike, and their numerical identity and spatial location is the only thing keeping them separate. Adding different kinds of parts is obviously a way of enriching the possibilities of organization. Hence our first dimension is:

(DO1) Diversity in the kinds of components

By diversity, we mean diversity in the intrinsic causal properties of the parts. As causal properties of the parts differ, the causal interactions that they can participate in also differ, thus making more variety possible in the causal activities of the system. An uneven distribution in the properties of parts is an elementary form of organization (an uneven distribution of pixels of different colours constitutes a picture), but just having diverse elements does not get one very far, though. Once the elements can be ordered and related in different ways, a much richer organization becomes possible. Hence, the second dimension is

(DO2) The relations between components (network structure)

An element of organization is introduced just by letting the relations between the parts be unevenly distributed. Uneven distribution in the relations between the parts amounts to the system having a particular network structure: the whole is not indifferent to which specific parts are related to which. Network structure is constitutively relevant to the property of the whole if changing the network structure would also change the property of the whole. The ‘new science of networks’ (Watts 2004) is precisely in the business of providing models which link different network structures to specific systemic properties, thus providing formal tools with which to answer such w-questions. Different properties of the network structure, such as small worlds or network modularity, and elementary network structures, such as the star, wheel or spanning tree, all have tractable repercussions on the behaviour of the whole that can be abstracted from the causal make-up of whatever system is realizing the network.

A further element of organization can be added to a system with components with varying properties and a particular network structure by allowing the relations between the component parts to have different properties. Hence the third dimension is

(DO3) Diversity in kinds of relations

Properties of a causal interaction between the parts include things such as duration, rate, inhibition, promotion, modulation etc. When the relations in the network are temporally ordered, the system has dynamic properties. The properties of relations explain system-level behaviour if we can link changes in the specific properties of such relations (say, a change from a linear to an exponential inhibition between properties of specific components) to specific changes in the system-level property.

A more and more complex organization can be added to a system by combining and iterating organizational features along the dimensions laid out above. However, we can understand an aspect of the behaviour of the whole with knowledge of the organization of its parts only when the consequences of possible changes in the organization (answers to w-questions) remain tractable for finite cognitive agents such as ourselves. Such inferences remain feasible when we first learn to reason with simple *organizational motifs*, abstract schemata that simultaneously combine only a very limited number of organizational features along the dimensions. Organizational motifs make it possible to make reliable w-inferences from changes in the motif to changes in the contribution that the motif has to the behaviour of the whole.

We will next demonstrate that our schema of the dimensions of organization follows research heuristics used in the study of complex systems and helps to make sense of the explanatory import such studies may have. We will use two examples: the computational models extending the original segregation model by Thomas Schelling and work on network motifs in transcription regulation networks controlling gene expression. The examples demonstrate how the use of organizational motifs facilitate understanding and how experiments and models aim at exploring the effects of changes in only a limited number of organizational aspects, preferring changes in only one dimension, at a time.

6 The Organization of Segregation

The checkerboard model of segregation by Thomas Schelling (1971, 1978) is one of the best known – and probably the most explored – agent-based simulation models in the social sciences (Fossett 2006; Aydinonat 2008). The model addresses the origins of residential segregation by race or ethnicity and is, due to its simplicity, a good platform to study the explanatory relevance of organization.

In the original two-dimensional checkerboard model, the model world consists of agents living on the squares of a checkerboard. The agents are divided into two classes that represent any binary social division that could affect the distribution of agents in space (e.g., blacks and whites, or humanists and engineers). While the model does not say anything about the origins of the agents' preferences, it assumes that each agent has a threshold for tolerating members of the other group in its neighbourhood. For example, an agent might prefer not to be in the minority or she might require that at least a third of the neighbours are from the same group. Initially, the agents are randomly distributed across the board, and some squares are left unoccupied. The agents can observe their Moore neighbourhood (the eight-cell combination of four adjacent cells and four diagonal cells), and they can change their location if the number of agents of the other type exceeds the threshold. When this happens, the dissatisfied agent randomly relocates to a new spot on the board.

One of the striking results of this model is that even when agents are highly tolerant of the opposite type, segregation is still likely to emerge. Segregation arises due to the phenomenon of tipping, whereby the early movements of even a few dissatisfied agents can create an incentive for others to move. This creates a cascade of movement that only dies out when the whole board has become highly segregated. This is the core feature of the model. Agents' attempts to avoid being in the minority by moving to a new location change the composition of both the old and new locations in a way that precipitates further movement. The neighbourhood that they leave becomes less attractive to members of their own group and the members of the other group find the neighbourhood that they enter less attractive after the move. Ultimately, over successive iterations, segregated neighbourhoods emerge.

What does this simple model tell about the explanatory relevance of organization? First, while this model is very simple, it still has a very interesting feature: the discontinuity between the properties of parts (the individual preference of avoiding being in minority) and the collective outcome (segregation). This discontinuity is of great social scientific interest, as it is a common fallacy to assume that segregation must be an outcome of discriminatory preferences. Second, while the model does not explain any specific instance of segregation – it is too abstract for that – it outlines a mechanism schema for a how-possibly explanation that has very robust results. It provides a general template for thinking about segregation processes far beyond residential segregation and the social sphere in general (Vinković and Kirman 2006; Clark and Fossett 2008).

The robustness of the model is apparent in two ways. First, most changes in the size of the neighbourhood, individual preferences or the availability of flats do not change the outcome. Second, even a small random move can lead a non-segregated area to a path that leads to segregation. This robustness as such is a signal that a specific organization does *not* matter (as long as the basic parameters are the same). However, the great advantage of agent-based computer simulation methodology is that it enables studying when and how organization begins to matter, i.e., the manipulation of individual organizational variables. First, we can manipulate what kinds of agents we have. For example, we can introduce heterogeneity in preferences or new attributes (such as wealth and social status) to the agents (Benard and Willer 2007). Thus, we can introduce diversity in the elements (DO1) at will. Second, we can manipulate the network structure (DO2). For example, we can change how the agents see their neighbourhood, or we can add irregularities into the spatial form of the neighbourhood. And third, we can make the rules more complex and thus introduce diversity to the relations between the agents (DO3).

When the complexity of the model increases along these three dimensions, the role of organization also increases. Thus it is possible to find the specific thresholds that break the robustness of the original organizational irrelevance. For example, it has been shown that segregation remains low as long as groups' preference targets do not exceed their population representations, that bounded districts increase segregation, that the form and size of the 'vision' of the agents influences the segregation pattern, etc. (Clark and Fossett 2008; Fossett and Warren 2005). Knowledge of such general dependencies enables counterfactual inferences concerning the system-level consequences of alternative types of the organization of parts. The great advantage of agent-based architecture is that it allows systematic and piecemeal study of these processes.

7 Network Motifs in Gene Regulation Networks

The transcription of proteins from genes is regulated by transcription factors that are sensitive to various signals – including the levels of other proteins and the level of the transcribed protein itself. These feedback and feedforward interactions dramatically expand the space of organizational possibilities and make it possible to create highly complex systems such as humans from a relatively limited number of genes. Experimental research and computational modelling on these patterns of activation and inhibition have revealed a limited number of recurring patterns, network motifs, with specific modular functional ('information processing') contributions to the behaviour of the whole. Such motifs are not limited to transcription regulation, but are also found in subsequent protein modification processes and networks between neurons (Mangan and Alon 2003; Alon 2007).

The most common simple motifs are feedback and feedforward loops composed of three factors in which the level of a transcripted protein (X) affects the transcription of another protein directly (Z) and through an intermediary (Y). All three of these links can be either promoting or inhibitory, leading to eight different motifs with different dynamic behaviours (Alon 2007). Also, the response of the

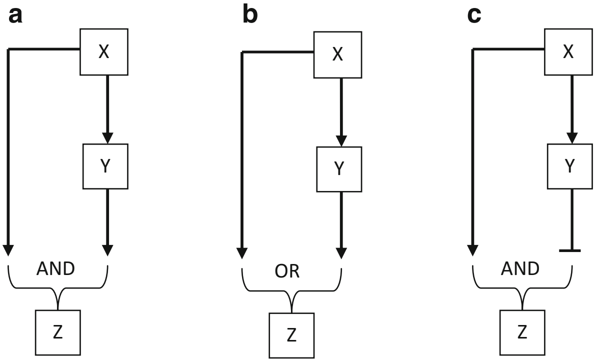


Fig. 1 Three common feedforward motifs

regulated factor can be of the AND- or OR-type (whether the transcription of Z is promoted/inhibited if both X and Y have reached suitable levels or if the presence of either one is sufficient), leading to further systematic differences in the dynamic behaviour and functional role of the motif. For example, in the most studied networks (E. Coli and yeast), the most frequent motifs are two such feedforward loops (FFL). The first is 'coherent' in that all the factors are promoting. The other is 'incoherent', in that the first protein (X) promotes the transcription of two others, of which one (Y) in turn inhibits the transcription of the other (Z). If the effect of X and Y on Z is of the AND-type, the coherent FFL motif (Fig. 1a) acts as a 'sign-sensitive' delay element in that there is a delay in the transcription of Z after X is turned on (since its transcription also requires the production of Y), but no delay in the negative regulation, since turning X off turns Z off almost immediately. With the OR-type functional dependence of the Z (Fig. 1b), the motif's role is reversed: the FFL shows no delay after stimulation of X, but does show a delay when the stimulation stops. The most common incoherent FFL (Fig. 1c) in turn acts as a pulse generator: the production of X causes the immediate production of Z, which is later turned off when the level of Y, also promoted by X, has accumulated to the level that effectively represses the transcription of Z.

Such simple organizational motifs with characteristic functional properties are combined in biological networks to produce greater functional complexity: a particular FFL can provide an input signal to another FFL and so on. The uncovering of this modular functionality of regulation networks facilitates understanding of the systemic behaviour in that we can now, in principle, answer w-questions concerning the consequences of local changes in the structure (DO2) and kinds of relations (DO3) in the whole regulation network by tracing the consequences of such a change through the sequence of network motifs according to their modular functional properties. What if the promotion of protein p had been regulated according to an OR-gate rather than an AND-gate? We can answer such questions by replacing the functional properties of the OR-gate motif with the functional properties of the corresponding AND-gate motif and tracing the system-level consequences of such a change. How would the dynamic behaviour of the network have been different if the transcription rate of a particular repressor in an incoherent FFL had been r’ rather than r”? Simulation studies and in vitro experiments have revealed how the dynamic behaviour of specific motifs is dependent on such changes in parameter values, thus enabling answers to such w-questions. Network motifs therefore constitute a paradigm example of the way in which iterated organizational motifs render such counterfactual questions tractable.

8 Conclusions

We have argued that the contrastive-counterfactual framework is able to fill an important gap in mechanistic theories of explanation: accounting for the role of organization in mechanistic explanations. This is a powerful argument in favour of the framework, since no other theory seems to provide similar tools. Organizational explanations trace constitutive dependencies between organizational motifs and system-level properties. Such organizational motifs in turn are characterized by differences along our dimensions of organizational dependence: diversity in the kinds of elements, the network structure and diversity in the kinds of relations. The whole is dependent on the organization of its parts in that if the motif had been different along one or more of the dimensions, the whole would have been different as well.

The importance of motifs shows how organizational explanation is facilitated by searching for ways in which the organization itself can be seen as composed of semi-independent ‘parts’. This raises the old chicken-and-the-egg question of whether we can understand much of the world because the architecture of complexity is usually suitably modular, or whether we selectively conceive only suitably modularly organized constellations of things as interesting objects of explanation. Whichever the answer, the search for modular organizational motifs is a powerful reductionist heuristic in the search for constitutive mechanistic explanations.

Acknowledgement We thank Caterina Marchionni and the audience at EPSA 2011 for their valuable comments. This research has been financially supported by the Academy of Finland.

References

Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics, 8*(6), 450–461.

Aydinonat, E. (2008). *The invisible hand in economics: How economists explain unintended social consequences*. London: Routledge.

Benard, S., & Willer, R. (2007). A wealth and status-based model of residential segregation. *The Journal of Mathematical Sociology, 31*(2), 149–174.

Clark, W. A. V., & Fossett, M. (2008). Understanding the social context of the Schelling segregation model. *Proceedings of the National Academy of Sciences, 105*(11), 4109.

Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York/Oxford: Clarendon.

Fossett, M. (2006). Ethnic preferences, social distance dynamics, and residential segregation: Theoretical explorations using simulation analysis. *The Journal of Mathematical Sociology, 30*(3–4), 185–273.

Fossett, M., & Warren, W. (2005). Overlooked implications of ethnic preferences for residential segregation in agent-based models. *Urban Studies, 42*, 1893–1917.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*, S342–S353.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*, 49–67.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7*, 368–373.

Mangan, S., & Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences, 100*(21), 11980–11985.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology, 1*, 143–186.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. London/New York: W. W. Norton.

Vinković, D., & Kirman, A. (2006). A physical analogue of the Schelling model. *Proceedings of the National Academy of Sciences, 103*(51), 19261–19265.

Watts, D. J. (2004). The “new” science of networks. *Annual Review of Sociology, 30*(1), 243–270.

Wimsatt, W. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge: Harvard University Press.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Ylikoski, P. (2009). The illusion of depth of understanding in science. In H. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 100–119). Pittsburgh: Pittsburgh University Press.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies, 148*, 201–219.

Mechanistic Explanation: Asymmetry Lost

Samuel Schindler

Abstract In a recent book and an article, Carl Craver construes the relations between different levels of a mechanism, which he also refers to as constitutive relations, in terms of means of mutual manipulability (MM). Interpreted metaphysically, MM implies that inter-level relations are symmetrical. But in that case MM violates one of the main desiderata of scientific explanation, namely explanatory asymmetry. Parts of Craver's writings suggest a metaphysical interpretation of MM, and Craver explicitly commits to constitutive relationships being symmetrical. Other parts of Craver's writings suggest an epistemological reading. If interpreted in this way, however, namely as a means for individuating mechanisms, MM is arguably redundant.

1 Introduction

Ever since Machamer et al. (2000)'s landmark article "Thinking about mechanisms", mechanistic explanations– thought to be the most pervasive kinds of explanation in the biological sciences – have become a major research topic in the philosophy of science. In a nutshell, Machamer et al. characterize mechanisms as being "composed of both entities (with their properties) and activities. Activities are the producers of change. Entities are the things that engage in activities" (p. 3). To provide a mechanistic explanation of a phenomenon, then is "*to explain how it was produced*" by a mechanism (ibid.). The *production* of the phenomenon in question by a mechanism, call it MPP (*m*echanistic *p*roduction of the explanandum *p*henomenon), is thus absolutely central to the mechanistic conception of explanation. Although

S. Schindler (✉)
Department of Physics and Astronomy, Centre for Science Studies, Aarhus University,
Munkegade 120, Building 1520, 8000 Aarhus, Denmark
e-mail: samuel.schindler@ivs.au.dk

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_7,

not explicitly highlighted by Machamer et al., MPP ensures that the mechanistic account of explanation captures one of the most important desiderata on accounts of explanation: explanatory asymmetry. Mechanisms explain phenomena, but phenomena do not explain mechanism, because mechanisms produce phenomena and not vice versa. The direction of explanation thus follows the direction of a mechanism's production of the relevant phenomenon. This assumption is in fact analogous to an assumption made by large parts of the philosophical literature on causation, perhaps most explicitly put by Salmon (1998, p. 129): "The asymmetry of explanation is inherited from the asymmetry of causation" (see also Strevens 2008b, 24f. and pp. 76–77). And indeed, although mostly concerned with the descriptive project of drawing to the attention of philosophers the importance of mechanistic explanations, Machamer et al. do express broad and general sympathy with a causal process theory for MPP in the tradition of Salmon's (1984) early work on causation. Process theories of causation, however, have widely been acknowledged to fail on various counts (Hitchcock 1995).

In his recent book (Craver 2007) and an article (Craver and Bechtel 2006), Craver offers important refinements of the original mechanistic account by Machamer et al. Amongst other things, Craver proposes that the relation between the mechanism and the explanandum phenomenon (i.e., MPP relation) be understood in terms of "mutual manipulability", which, a reviewer has judged to be "one of the main achievements of the book“ (Levy 2009, p. 141). It will be the purpose of this paper, to assess this aspect of Craver's account.

This is how I proceed. In Sect. 2 I introduce Craver's notion of mutual manipulability (MM) as an explication of MPP. I argue that Craver's explication of MPP strips the mechanistic account of explanation of its ability to capture explanatory asymmetry. In Sect. 3 I explore ways in which this undesirable consequence might be avoided. One option I highlight is the interpretation of MM as a purely epistemological criterion for identifying MPP's. As I argue in Sect. 4, however, there is clear textual evidence that Craver intends MM as an explication of the *meaning* of MPP, which is a genuinely metaphysical project. If interpreted in purely epistemological terms, I argue in Sect. 5, MM becomes redundant. In Sect. 6 I conclude this paper by recommending the abandonment of MM and by pointing to one feature of mechanisms that the proponents of the mechanistic approach might want to focus on in order to justify the need for account of *mechanistic* explanations.

2 Explanatory Asymmetry Lost?

Just like Machamer et al. (2000), Craver (2007, pp. 6–7) defines mechanisms as "entities and activities organized such that they exhibit the *explanandum* phenomenon". Craver's (and Machamer et al.'s) standard example for a mechanistic explanation is the explanation of the neuronal action potential, which "is explained by reference to component parts of the action potential mechanism", whereby examples for component entities are ions, ion channels, protein chains, etc. and examples for component activities are diffusion processes, and changes in confirmation (pp. 121–122).

Craver distinguishes between a lower and an upper level in mechanisms (pp. 6–7). At the lower level he locates the entities X and their properties or activities ϕ; the 'upper' level is constituted by the phenomenon to be explained. The 'mechanism as a whole', i.e., X, ϕ, and the explanandum phenomenon, Craver denotes with S. Furthermore he treats the phenomenon to be explained as being equivalent to S's activity ψ. MPP then, in Craver's terminology, is S's ψ-ing (i.e. the explanandum phenomenon) being "exhibited" or "produced" (Craver uses both terms) by the activities of the mechanism's components (X's ϕ-ing). Furthermore, even though of minor importance in the following, for Craver (as for Machamer et al.) mechanisms often consist of multiple levels. That is, the upper level of one mechanism may be a component of a lower level of another mechanism, and so on. More importantly, Craver sharply distinguishes between intra-level and inter-level relations (Craver and Bechtel 2006; Craver 2007). Whereas intra-level relations are causal relations, inter-level relations are not; they are so-called *constitutive* relations. Constitutive relations – in contrast to causal relations – are symmetric, synchronous, and part-whole relations (pp. 153–154). Although not made very explicit by Craver, constitutive relationships are meant to specify MPP, as we shall see in the following.

In order to elucidate inter-level relationships in mechanisms, Craver (2007), following Woodward (2003), adopts the notion of an ideal intervention: "an ideal intervention I on ϕ with respect to ψ is a change in the value of ψ that changes ψ, if at all, only via the change in ϕ" (pp. 154). Interventions need not be performable by humans, nor need they be physically possible. All that is required is that they be logically possible (see Woodward 2003, 127ff.). Craver (2007)'s explication of inter-level relations in terms of ideal interventions consists of two parts, which together form his *mutual manipulability* criterion (MM):

(CR1): When ϕ is set to the value ϕ_1 in an ideal intervention, then ψ takes on the value of $f(\phi_1)$. (p. 155)
(CR2): When ψ is set to the value ψ_1 in an ideal intervention, then ϕ takes on the value of $f(\psi_1)$. (p. 159)

Apparently, both CR1 and CR2 have the structure of Woodwardian active counterfactuals, i.e., counterfactuals whose antecedents are made true by interventions which Woodward intends to pick out *causal* relationships. And yet, Craver denies that neither CR1 nor CR2 do so. As mentioned above, the combination of CR1 and CR2 (i.e. MM) is supposed to individuate *constitutive* relations, which, according to Craver, are not causal relations. In accordance with the convention in the contemporary literature on causation to refer to a causal relation between X and Y as X "being causally relevant" to Y (cf. Woodward 2003, p. 39), Craver also refers to CR1 and CR2 as criteria for constitutive *relevance.* More specifically, "one can change the *explanandum phenomenon* by intervening to change a component [of a mechanism]", *and* vice versa, "one can manipulate the component by intervening to change the *explanandum phenomenon*" (p. 153). Craver concludes that

> many, if not most, causal relationships are unidirectional. In contrast, all constitutive dependency relationships are bidirectional (p. 153).

And since inter-level relationships are symmetrical relationships, they are therefore "only uncomfortably viewed as causal" (p. 153). Another reason that Craver mentions for constitutive relationships being distinct from causal relationships, which I will not be concerned with in much detail here, is that in the former, "ϕ's taking on a particular value is not temporally prior to ψ's taking on its value" (pp. 151–152), in other words, constitutive relationships are 'synchronic', whereas causal relationships are not. However, contrary to what Craver seems to suggest, a relationship being symmetrical does not imply that it cannot be causal (cf. Woodward 2003, p. 396). Furthermore, as Leuridan (2012, fn. 27 and 29) points out, many relations of interest in neurobiology (Craver's subject) are causal feedback loops, i.e., symmetrical causal relationships.

Craver's explicit commitment to inter-level relationships being symmetrical relationships raises the following concern: if inter-level relationships really are symmetrical, what is it in Craver's account that ensures that the desideratum of explanatory asymmetry is respected? Recall, on the original mechanistic account by Machamer et al., explanatory asymmetry is respected, because mechanisms *produce* phenomena, but not vice versa. The direction of explanation simply follows the direction of production. So what happened to the production relationship in Craver's account, which I referred to as MPP, and which Craver, like Machamer et al., characterizes as a central feature of mechanisms? At one point in his book, Craver *seems* to say that MPP is to be spelled out in terms of Woodwardian counterfactuals picking out *causal* relations:

> *To say* that one stage of a mechanism is *productive* of another (as I suggest in Machamer et al. 2000; Craver and Darden 2001) is *to say*, at least in part, that one has the ability to manipulate one item by intervening to change another. (pp. 93–94; added emphasis)

As mentioned above, however, for Craver inter-level relations are *not* causal relations. So if the above quotation *were* to refer to MPP (i.e. an inter-level relation) Craver would clearly contradict himself. So despite speaking of "production" in this one passage, suggesting reference to MPP, Craver is better read as spelling out *intra*-level causal relations here. This still leaves us with the question of how MPP is to be understood.

In personal communication,[1] Craver is ready to give up on a 'literal' reading of MPP; he suggests MPP be interpreted metaphorically instead. That is, whenever we say that the explanandum phenomenon is "produced" by the mechanism, what we should say more carefully is that the phenomenon is constituted or "made up"

[1] Email communication with the author of this paper on 02-22-2012.

by the mechanism, very much in accordance with a constitutive understanding of inter-level relations. This is in line with the one taken in Craver and Bechtel (2006):

> The causal claims, when made explicit, are all intra-level. But we continue to talk about bottom up causal relation [from one to another level] when we are being quick or informal as long as we understand that the change at the higher level is mediated by, or explicable in terms of, a mechanism [and its constitutive relations]. (ibid, p. 557)

In this context it is furthermore interesting to note that the idea of MPP, i.e., the idea of a mechanism's producing the explanandum phenomenon, is altogether absent from Craver and Bechtel's joint paper. But again, construing MPP as a symmetrical relation (namely as CR) makes the mechanistic account vulnerable to the problem of explanatory symmetry. If the relation between the mechanism and the phenomenon is symmetrical, what is it that prevents us from saying that the phenomenon also explains the mechanism? In response to this question, Craver (personal communication) is ready to embrace a deflationary "explanatory pluralism", implying that phenomena might as well explain mechanisms. But perhaps there are more appealing options for Craver.

3 Explanatory Asymmetry Saved?

There is a much simpler response to the problem of explanatory symmetry available to Craver. He could point out that MM is only part of how constitutive relationships are to be understood. Another important aspect of constitutive relationships is that they are *part-whole* relationships (see above). Part-whole relationships are asymmetrical relationships: if ϕ is a part of ψ then ψ cannot be part of ϕ. Explanatory asymmetry between ϕ and ψ would thus be secured.[2] The problem with this response, however, is that it stands in outright contradiction with Craver's assertion that constitutive relationships are symmetrical: either constitutive relations are part-whole relations or they are symmetrical. Craver cannot have it both ways. Although the former option seems much more plausible in the face of the problem of explanatory asymmetry, part-whole relations are underdeveloped in Craver's account. In fact Leuridan (2012) argues that under the perhaps most intuitive definition of part-hood, cases of mutual causation cannot be ruled out by CR. This is contrary to Craver's insistence that constitutive relations are different in kind from causal relations. Furthermore, if Craver were to drop the idea of constitutive relationships being symmetrical, he would rob himself of one of the main characteristics distinguishing constitutive and causal relations (Craver and Bechtel 2006; Craver 2007). At any rate, there appear to be only two options for Craver: either he gives up on the idea that constitutive relations are symmetrical or he tries to save explanatory asymmetry whilst holding onto constitutive relationships

[2] Craver did not make this reply in the abovementioned email correspondence. I thank an anonymous referee for making me consider this option.

being symmetrical (and giving up on part-hood). The former option appears much more plausible. It would allow Craver to embrace the asymmetry property of part-whole relations and thereby explanatory asymmetry. But because Craver is so adamant about constitutive relations being symmetrical, let us briefly consider the latter option, before exploring the former option.

Take one of Woodward's preferred examples, the ideal gas law. This law relates variables of pressure (P), the volume of a gas (V) and temperature (T) in the formula PV=RT (R is the gas constant). Clearly this is a symmetrical relationship. We can intervene on P to change T, and conversely, we can intervene on T to change P (whilst holding fixed V in both cases). Now, assuming (with Woodward) that the ideal gas law is an explanatory generalization, in each of the above scenario, explanatory asymmetry is preserved despite the relationship being a symmetric relationship. In the one context T is the cause variable and P the effect variable, and in another context, P is the cause variable and T the effect variable. In the first scenario a change in T explains the change in P (but not vice versa), and in the second scenario P explains T (but not vice versa).[3] There is of course no a priori reason why this insight could not be extrapolated to the context of mechanistic explanations. However this extrapolation is only of a limited sort. It extends only to intra-level but not to inter-level relationships in mechanisms. That is, on one particular level of a mechanism it might make sense to say, as in the above example, that a change in X_1's ϕ_1-ing explains (in a minimal sense) a change in X_2's ϕ_2-ing, and vice versa (just in case, of course, X_1 and X_2 are related as P and T above). Further, it also makes sense (now between different levels of mechanisms) to say that if we can intervene on ϕ's to change ψ (the explanandum phenomenon), then ϕ's explain ψ. However, crucially, it makes little or no sense to say the reverse, namely that ψ explains ϕ, *even if CR2 were satisfied.* After all, the explanandum phenomenon cannot explain the mechanism, at least not in the standard sense of the term. It therefore seems that there is no straightforward way in which Craver could stick to his symmetry thesis and save explanatory asymmetry.

As mentioned above Craver could simply give up on his symmetry thesis. Indeed, MM implies the symmetry of inter-level relations *only if* it is to be understood as a *metaphysical* explication of inter-level relations, i.e., as specifying the meaning of inter-level relations. If read in epistemological terms, that is, as a criterion for *identifying* inter-level relations (in contrast to, say, mere correlations between a mechanism and some phenomenon), MM has no implications for the directionality of inter-level relations. At least a priori, there is no contradiction between inter-level relations being asymmetrical and our means for identifying them being applicable in both directions of this relation (i.e., bottom-up *and* top-down). An epistemological interpretation of MM would also allow Craver to reconcile MM with the asymmetry of part-hood. Unfortunately for Craver, however, there is strong textual evidence that Craver aims for a metaphysical explication of constitutive relevance in terms of MM. So before we can consider a pure epistemological interpretation of MM, we need to consider that textual evidence in more detail.

[3] This was suggested to me by Bert Leuridan.

4 The Metaphysics of Mechanisms

First and foremost, as pointed out above, part of Craver's motivation to amend the original mechanistic account by Machamer et al. is clearly owed to the wish to explicate the *meaning* of MPP. Such endeavor is generally regarded as a genuinely metaphysical. Second, Craver's concession that MM implies the symmetry of inter-level relations clearly presupposes that MM is interpreted metaphysically. Again, if MM were a mere epistemological criterion, nothing would follow for the direction of inter-level relations. Furthermore, Craver makes clear that he wishes to provide a 'normative' account that can "demarcate [mechanistic] explanation from other kinds of scientific achievements", and that can "reveal criteria for *assessing* explanations", in other words it "should prescribe norms of explanation" (p. 20; original emphasis). All this he could not do if he were to interpret MM epistemologically rather than metaphysically. One e.g. cannot assess whether an explanation is a good explanation without having provided at least a partial answer to the question of what constitutes an explanation, i.e., a question about the meaning of explanation.

But again, there are indeed also a number of passages in Craver's book that suggest that Craver views MM as an epistemological criterion. Primarily, this is suggested by the context in which Craver explicates inter-level relations. This context is formed by Craver's pointing to the various inter-level experimental strategies that can be used, inter-level bottom-up and top-down, to *establish* certain entities and activities as being part of a certain mechanism. Verbatim, Craver says:

> I build my positive account [of mechanisms] by considering the *experimental* strategies that neuroscientists *use to test* whether a given entity, activity, property, or organizational feature is relevant to the behavior of the mechanism as a whole [. . .] (p. 140; added emphasis).[4]

Furthermore, Craver appears to think that a non-reductive analysis of causation (or constitution) based on active counterfactuals implies that the analysis is non-metaphysical. Explicitly, he says in his summary of Woodward's manipulationist account, which he then goes on to use to define constitutive relevance in mechanisms, that

> I do not discuss here whether such metaphysics [of causation] is required or what the available metaphysical options are. Even if the manipulationist view does not identify the truth-maker[s] for causal claims, it is nonetheless an illuminating analysis of the causal truths themselves [. . .] (Craver 2007, pp. 105–107).

But this is a misapprehension. The fact that a philosophical analysis is non-reductive does not imply that it is non-metaphysical. To see this requires a short excursion into Woodward's account.

Many philosophical analyses seek to reduce the concept of causation to another concept. Humeans, for instance, reduce causation to mere empirical regularities.

[4] Couch (2011, fn. 6) also reports that Craver explicitly embraces an epistemological interpretation of MM in personal communication.

David Lewis reduces causation to counterfactual dependence. Woodward's analysis of causation, in contrast, is decidedly non-reductive. Woodward defines causal relationships as generalizations that remain invariant under interventions. As Woodward acknowledges, the notion of an intervention is itself a causal notion. In a sense, Woodward's account is thus circular. However it is not *viciously* circular, as Woodward (2003, p. 20ff.) points out, because the causal relation that is being appealed to (*I* causing a change in *X*) is different from the causal relation that the analysis seeks to illuminate (namely, X causes Y). Woodward goes on to show that such a non-reductive account can very well be insightful. It for instance delivers markedly different verdicts on a number of important issues (e.g. action-at-a-distance, causation by prevention or absence) than Salmon (1984)'s classical causal process theory of causation.

The crucial question now is of course: is a non-reductive analysis of causation non-metaphysical (as Craver would have it)? This is not an uncontroversial matter. In a slightly heated exchange with Strevens (2007, 2008a), Woodward (2008) explicitly denies that his own account is metaphysical. However Strevens (2008a) offers a number of convincing reasons why Woodward might in fact be mistaken about the aims of his own book. First, Woodward (2003), throughout his book, presents his analysis as a superior rival to Lewis's metaphysical analysis of causation. It would be hard to see why Woodward would do so, if the aim of his project were entirely different from Lewis's. Second, Woodward seeks to provide an account according to which causation is mind-independent, in other words, an account of the *nature* of causation which is independent of how we get to know about this relation (2003, p. 118ff.). This clearly is a metaphysical endeavor. Third, Woodward states that "my aim is to provide an account of the meaning or content of various locutions, such as X causes Y" and that "my project is semantic or interpretive" (p. 38). As Strevens (2008a) points out:

> In modern times, such a project is invariably interpreted as aiming to provide truth conditions for the sentences or thoughts in question, and therefore as aiming to specify those representations' truthmakers. It may look like semantics, but it is also a kind of metaphysics [. . .] it is generally agreed that a word with an explicit definition has as its extension whatever stuff satisfies that definition. If Woodward's causal semantics is a truth-conditional semantics, he is inevitably, unavoidably, ineluctably committed to producing an account of the truthmakers for causal talk, a metaphysics of causal facts, whatever his protestations. (p. 184)

Back to Craver. If Craver wishes to spell out the *meaning* of the inter-level relation in terms of constitution then, by Strevens's lights, he inextricably commits himself to a metaphysical project. Since there are clear signs that Craver does wish to elucidate the meaning of inter-level relations (see above),[5] his project, contrary to what he says himself, does commit him to a metaphysics of constitutive relations. And since Craver's spelling out of inter-level relations implies a symmetrical inter-level relation, which he explicitly embraces, the mechanistic account no longer

[5] This is also what Craver told me in personal communication.

captures explanatory asymmetry. So let us now finally consider the possibility of interpreting MM, contrary to much of what (is implied by) what Craver says about it, as a merely epistemological tool for individuating constitutive relationships.

5 Individuating Mechanism Boundaries

The main epistemological function Craver assigns to MM, which I alluded to briefly above, is this: it concerns the delineation of the *boundaries* of mechanisms (Craver 2007, p. 141ff.). In other words MM is supposed to specify (i) which entities and activities are, and which ones are not, part of the lower level of a mechanism with respect to a particular explanandum phenomenon (this function is performed by CR2), and, conversely, (ii) which phenomenon is picked out by a particular mechanism (this function is performed by CR1). Craver gives the following example (p. 152).

The cognitive capacity of word-stem completion, in which a subject is presented with a list of words and afterwards asked to complete the word stems of the words presented previously, is affected by changing the heart rate of the subject. That is, if one were to change the heart rate of the subject (by e.g. torturing the subject), the subject's capacity to complete word stems would invariably change. According to CR1 *alone*, however, the heart rate would *erroneously* be deemed as constitutively relevant for the phenomenon of word-stem completion; the heart rate normally would not be considered a part of the mechanism of word stem completion, or so Craver reasons. This is where CR2 comes in. Engaging subjects in word-stem completion will *not*, under normal circumstances, result in a different heart rate. Hence CR2 is not satisfied by the example and the heart rate is therefore not to be deemed part of the mechanism of word stem completion.

It is questionable whether MM really fulfills the purpose Craver allots to it. To see this, note that the persuasiveness of the above example rests on the implicit assumption that the relevant mechanism for the capacity of word stem completion is a *cognitive* mechanism. Only then the heart rate appears irrelevant. But not in all contexts need this be so. In contexts in which one is interested in, for instance, the capacity of completing word stems as compared to the absence of any cognitive capacity, the heart rate appears to be indeed a part of the relevant mechanism. In other words, whether or not a mechanism (or part thereof) is relevant to the explanation of a phenomenon is subject to pragmatic considerations. In fact Craver is very well aware of this. He for instance highlights the importance of contrast classes in the specification of the explanandum phenomenon (2007, p. 202ff.).

> [. . .] the spatial and causal boundaries of mechanisms depend on the epistemologically prior delineation of relevance boundaries. But relevance to what? The answer is: relevance to the phenomena that we seek to predict, explain, and control. Within the boundaries of a mechanism are *all and only* the entities, activities, and organizational features relevant to the phenomenon selected as our explanatory, predictive, or instrumental focus. (Craver 2009, p. 591; added emphasis)

The epistemological function CR2 is supposed to perform, namely the picking out of a particular mechanism, *given a particular explanandum phenomenon*, is therefore not necessary. It is plausibly carried out by pragmatic considerations: we choose a particular phenomenon we want to explain, predict, etc. and then we ask, by reverse engineering, as it were, what causes are responsible for the phenomenon in question. MM as an epistemological criterion in Craver's account looks therefore redundant.

6 Conclusion

When interpreted metaphysically, MM, in violation of a central desideratum of explanation, implies explanatory symmetry and is inconsistent with the part-hood characterization of constitutive relevance (as I argued in Sects. 2 and 3). On the other hand, when interpreted epistemologically, MM does not give us any extra purchase on the individuation of mechanisms (as argued in the last section). I therefore believe that MM is better to be abandoned altogether. I thus disagree with Leuridan (2012) who concludes his detailed discussion of MM by suggesting that mechanistic inter-level relations be kept and interpreted, contrary to Craver, as relations of *mutual causation*. But such a proposal is of course just as much subject to my criticism of explanatory asymmetry being lost (see also Sect. 3). Rather I think that inter-level the productive relation MPP is perhaps best understood in terms of *unidirectional* Woodwardian counterfactuals picking out *causal* relations, without invoking top-down counterfactuals (contra Craver). But of course, there would then be no need for a specific mechanistic account of explanation (see Woodward 2002). So is there?

A crucial feature of mechanisms is the *organization* of entities and their spatio-temporally *concerted* interaction that produces a phenomenon. In order to accommodate this feature of mechanisms, Woodward (2011) suggests that counterfactuals describing causal relations in mechanisms possess "characteristic spatio-temporal signatures" (p. 427). I take it that this suggestion translates into mechanisms being specifiable in terms of conjunctions of active counterfactuals with complex antecedents of the following form: <If an appropriate intervention *I* had changed the value of an "entity variable" X at time t1 *and*, if an appropriate intervention *I* had changed the value of an "entity variable" Y at time t2, *and* . . . etc., then the explanandum phenomenon would have been produced by the mechanism comprising entities X, Y, etc.>. Spatial location might be transcribable in terms of variables in a similar way. So barring concerns about the modularity assumption that Woodward makes (Cartwright 2002), such an amendment of Woodwardian counterfactuals to accommodate the genuine mechanistic feature of organization and concerted interaction of mechanism-components might be a fruitful way of cashing out mechanistic explanations without the need for a specific account of mechanistic explanation.

Acknowledgements This article has profited significantly from discussions with Bert Leuridan, Kareem Khalifa, Petri Ylikoski, Marcel Weber, and Alexander Reutlinger. An anonymous referee of this journal provided helpful comments. Carl Craver kindly clarified his views for me in personal communication. I'm grateful to all of these individuals. I furthermore want to thank the audience at the EPSA 2011 conference in Athens (Greece) for feedback.

References

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science, 53*(3), 411–453.

Couch, M. B. (2011). Mechanisms and constitutive relevance. *Synthese, 183*(3), 375–388.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology, 22*(5), 575–594.

Craver, C. F., & Bechtel, W. (2006). Top-down causation without top-down causes. *Biology and Philosophy, 22*(4), 547–563.

Hitchcock, C. R. (1995). Salmon on explanatory relevance. *Philosophy of Science, 62*(2), 304–320.

Leuridan, B. (2012). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *The British Journal for the Philosophy of Science, 63*(2), 399–427.

Levy, A. (2009). Carl F. Craver, Explaining what? Review of explaining the brain: Mechanisms and the mosaic unity of neuroscience. *Biology and Philosophy, 24*(1), 137–145.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. (1998). *Causality and explanation*. New York: Oxford University Press.

Strevens, M. (2007). Review of Woodward, making things happen. *Philosophy and Phenomenological Research, 74*(1), 233–249.

Strevens, M. (2008a). Comments on Woodward, making things happen. *Philosophy and Phenomenological Research, 77*(1), 171–192.

Strevens, M. (2008b). *Depth: An account of scientific explanation*. Cambridge: Harvard University Press.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science, 69*(3), 366–377.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research, 77*(1), 193–212.

Woodward, J. (2011). Mechanisms revisited. *Synthese, 183*, 409–427.

Deflationism on Scientific Representation

Chuang Liu

Abstract This paper critically discusses a deflationary view of scientific representation, which sees models in science and technology as no different in essence from other sorts of representational vehicles and regards all of them as derivative devices determined by convention. To reject the view, it is first argued that there are at least two radically different roles that representation plays, one is purely symbolic and therefore conventional, and the other is epistemic. The failure to recognize the epistemic role of representation, which is the main role for models in science and technology, led to the mistaken view that models are just like other symbols, such as the linguistic ones, and that defationism is the right answer to the constitutional question of scientific representation. The paper briefly considers in passing some broader questions in connection with the criticism of deflationism.

1 The Deflationary View

To the questions concerning the nature of scientific representation, which includes theorizing, modeling, and other activities and their products, numerous philosophical inquires have formed a sizable literature in recent years (cf. Morgan and Morrison 1999; van Fraassen 2008, and references therein). A recent work

C. Liu (✉)
Center for Philosophy of Science and Technology, Shanxi University (visiting), Taiyuan, China

Department of Philosophy, University of Florida, 330 Griffin-Floyd Hall, P.O. Box 118545,
Gainesville, FL, USA
e-mail: cliu@phil.ufl.edu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_8,

(Callender and Cohen 2006), defending a deflationary view, promises to cut the Gordian Knot and close the book on all the controversies, and this paper examines the view to point out its shortcomings.

Callender and Cohen begin their argument by pointing out that much confusion in the literature comes from trying to provide answers for the wrong questions: a case in point: people have been trying to figure out in what sense a model could be said to resemble – in terms of similarity or isomorphism – its target while addressing the question of what *constitutes* the relationship between the two. This mistake, they argue, is caused by confusing the 'constitution question' about models or modeling with such questions as the 'demarcation problem' (whose solution demands some sort of criterion for distinguishing those representational devices that can from those that cannot serve specific purposes in science and technology) and the 'explanatory/normative problem' of scientific models and modeling, which is about such questions as in virtue of what may scientists tell correct or explanatorily superior models from incorrect or explanatorily inferior ones.

These distinctions are long overdue, and one couldn't help but recalling a similar situation in the search for a theory of truth. Amid the controversy over which of the alternative theories, correspondence, coherence, or pragmatic, holds the truth on truth, Tarski's disquotational scheme, i.e. the scheme tokened by, for instance, 'Snow is white' is true if and only if snow is white, gives rise the hope of a deflationary theory, which views 'is true' either as a *redundant* predicate (a la Ramsey) or as a syntactic symbol for disquotation (a la Quine) or . . . . If there are deeper questions about how humans are able to produce and judge true statements, such as those legitimate questions that the correspondence theorists or pragmatic theorists ponder, they are not about the truth predicate (cf. Burgess and Burgess 2011).

To argue for deflationism, Callender and Cohen point out that scientific representation should naturally be regarded as a species of representation in general and what philosophers of language, such as Grice, has worked out for representation in general should also apply to scientific representation. Grice's theory of speaker meaning/representation – or what they refer to as the 'Specific Griceanism' – is a reductive account of how a speaker's utterances get their meaning from the conveying of the speaker's meaning-intentions. In other words, X means that p by uttering s if and only if X intends that the listeners of s forms the belief that p. The mental/belief states so invoked are the 'fundamental representations,' while the words (or other vehicles) that are used to invoke them are the 'derivative representations.' The latter represent via the former or in virtue of the former.[1]

Scientific representational devices, such as models, do their job in accordance with General Griceanism, which is a natural extension of Specific Griceanism. The basic scheme is the same, and it gives a unified account of how any derivative

[1] Cf. three seminal articles by Grice: "Meaning," "Utterer's Meaning, Sentence Meaning, and Word-Meaning," and "Utterer's Meaning and Intentions," all in Grice (1989).

representational devices do their job in representing the world to us. To illustrate their point, Callender and Cohen mention such acts of representation as lanterns being raised in a certain way at a certain hour to represent the presence or absence of enemy troops, or more dramatically, salt shaker on your dinner table being used to represent your favorite geographical region, e.g. Madagascar (Callender and Cohen 2006, pp. 13–14). The key and only condition of adequacy is that the right belief states are *intended* and *invoked* among the users of the devices, and for that, no other constraints, such as resemblance or similarity between the devices and the targets, are necessary. Since the success of a representational attempt has nothing to do with either the intrinsic or relational properties of the representing and the represented, anything can be used to represent anything else.[2]

Be that as it may, Callender and Cohen are by no means dismissive of the earlier efforts about the nature of scientific models and modeling. But they believe that *the answers to all the other questions, such as the demarcation and the explanatory/normative, are exclusively questions about the pragmatics of device usage*.

To summarize: (i) constitution question has been confused with other questions; (ii) constitution question admits a deflationary answer, and (iii) all the other questions only admit answers of a pragmatic sort. By General Griceanism, how we represent the world around us is reduced to (1) how conventionally selected external vehicles are related to the beliefs states, and (2) how those belief states represent.

2 Analyzing Deflationism

If the constitution question of scientific representation is construed as 'what can be used as a vehicle to represent a target under scientific or technological research?' deflationism must be correct. In other words, the question is construed (by Callender and Cohen) in such a way that it is about no more than the constraints of symbol-using by cognitive agents. Though the business cannot be entirely constraint-free, the constraints are of a rather trivial kind, some of which are sampled below.

[2] In this respect, Teller (2001) should also be regarded as a deflationist, especially when he says,

> I take the stand that, in principle, anything can be a model, and that what makes a thing a model is the fact that it is regarded or used as a representation of something by the model users. Thus in saying what a model is the weight is shifted to the problem of understanding the nature of representation (Teller 2001, p. 397).

Here, the talk of being 'regarded' or 'used' as a representation clearly implies that what makes something a model depends exclusive on a stipulation/convention in the community of model users; and 'the problem of understanding the nature of representation' clearly concerns the fundamental or natural representations. And so, perhaps to a lesser degree, is van Fraassen. He observes that if one is to have a theory of representation (which he doesn't) one must accept what he takes to be the '*Hauptsatz*': "*There is no representation except in the sense that some things are used, made, or taken, to represent some things as thus or so*." (van Fraassen 2008, p. 23).

Something has to be perceivable by its users to be eligible for playing the role of representation. Invisibles or inaudibles cannot serve the purpose. And, if A is to be used to represent B, A must not have a lesser degree of usability than B has, where the term ‘usability’ should be understood in a broad pragmatic sense. What pragmatic concerns could possibly make a community use, to reverse Callender & Cohen’s own example, Madagascar to represent the salt shaker on a dinner table or sick people in huts to represent the hanging of animal skins above the entrances? This is especially true between words and their referents, e.g. ‘cat’ can represent cats but not vice versa. And these constraints, trivial though as they are, in fact show that the (constitution) question of scientific representation is here taken to be entirely about *a pragmatic matter of convention*, where by “convention” I do not mean, as its narrow meaning may suggest, “by explicit agreement;” I mean “conventional in principle,” which is compatible with long established habitual agreements in a community, which may well be naturally shaped.

Now let us see how exactly a representational device succeeds in representing its target according to deflationism. Take an example in which hanging some animal skin above the entrance indicates that somebody is sick inside.[3] A user of such a symbol clearly intends to communicate something and she succeeds when the other members of the community realize what is intended. The use of the symbol invokes by convention the appropriate belief state in others about the existence of a sick person inside that entrance. That belief state that gets called up by the perception of the symbol, the derivative device, is the fundamental device that constitutes the basis of representation. Without that belief state populated in people’s heads, the animal skin above an entrance can play no representational role in the community.

What such belief states are and how they function to fulfill their representational role are questions, the search for whose answers filled the history of philosophy, especially in epistemology and philosophy of mind. For Descartes and Locke – I shall call such philosophers *representationists* – reality is represented to us via ideas (or percepts in Russell’s language) in our mind that serve as the primary representational devices,[4] but for Reid (and his followers, whom I shall call

[3] Callender and Cohen uses an example of an upturned right hand representing the state of Michigan to explain how General Griceanism works and said:

> [I]n each case, the story is that the left hand represents what it does (a cat, a fact about a cat) by virtue of (i) an analogous representational relation that obtains between a mental state and its object (alternatively, a cat or a fact about a cat), together with (ii) a stipulation that confers upon the left hand the representational properties of that mental state. Indeed, the easy adaptability of the Gricean story to these different sorts of representation is a mere corollary of its indifference to the kinds of things that serve as representational relata. (Callender and Cohen 2006, p. 14)

[4] The belief states don’t have to be ‘iconic’ or ‘pictorial,’ but they must have representational content. Despite the suggestive examples used in this paper, there is no suggestion that all belief states that serve as the fundamental representations must be iconic. For how we represent the world through perception, see Freeman (1991) and Siegel (2011).

non-representationists), no such devices exist.[5] Does Callender & Cohen's reductive account, which has its origin in Specific Griceanism, assume representationism? *Prima facie*, it does, because the reductive base is supposed to comprise belief states that supply 'fundamental/natural' representations. If so, a detailed answer to how the hanging of a piece of animal skin represents a sick person inside a hut would presumably, first, give an account of how the belief state that identifies that state works as a device in each mind to represent *the sick person in the hut*, and then give a separate account of how by convention animal skin above an entrance is used to signal and communicate the presence of a sick person by causing the appearance of that natural device in the appropriate heads. *Both accounts are necessary for a full account of how humans represent the world around them.*

It would be difficult to conceive a non-representationist alternative on this Griceanist scheme. First, where do we place the fundamental/natural representations that are mental states? Even if we modify it or give it up, it is still difficult to see how an external device, such as the animal skin in our case, is related to its target (e.g. the sick person) through some mental states that do not themselves represent. One may come up with a dispositional account of the fundamental devices in people's mind. Instead of thinking that what is invoked in people's mind by seeing a hanging animal skin is some belief state called out from memory whose content has a sick person in a confined space, a non-representationist may say, for instance, that the invoked state is some kind of dispositional state of the mind that is ready to give report on the conditions of a sick person in a particular confined space. Whether or not this alternative can work, it does not seem to be something the deflationists are ready to embrace.[6]

3 Criticizing Deflationism

Deflationism, as I said, should hold if it is a question of the necessary condition for using one thing as a symbol for another. But is this the only way to understand the constitution question of scientific representation, as Callender and Cohen has urged? I think the answer is 'No.' This is so because, *first*, it goes against our experience with the practice of model-building in science and technology to say that most questions scientists confront there are no more than essentially concerning "which conventional devices are the most practical to use." And *secondly*, there is a deeper reason: some representational vehicles are primarily pragmatic and determined by

[5] At least, that is not the case when one is directly perceiving what is in front of her (when we see a table, we don't see the image of it in our head while that image is connected to the table in one way or another such that our representation of it may be veridical in one case but illusory in another).

[6] The discussion here about representationism and its opposite is not meant to stand on its own (for which a survey of the contemporary literature is pre-requisite). Primitive though it is, it is intended to flesh out some *possible* details of deflationism that follow General Griceanism.

convention, some are primarily epistemic and determined by their epistemic virtue, and some are a mixture of both, namely, convenient symbols that also exhibits epistemic virtue. I now argue for these two points in turn.

When asked what deflationism can say about most of the significant questions concerning scientific representation, Callender and Cohen say the following.

> But note that, just as in the case of similar questions about non-scientific representations, the questions about the utility of these representational vehicles are questions about *the pragmatics* of things that are representational vehicles, not questions about their representational status per se. Thus, if the drawing or the upturned right hand should happen not to rank highly along the dimensions of value considered so far, this would, on our view, make them non-useful vehicles that do represent, rather than debar them from serving as representational vehicles altogether. (Callender and Cohen 2006, p. 15, my italics)

If apart from some trivial constraints mentioned above anything can be used as a model for anything else in science, scientific models would resemble currencies an economy adopts for its economic transactions. The constitution question about what can be used as money is indeed entirely a matter of pragmatics, of convenience and utility. Do models or representation vehicles really resemble money in the relevant respects? Well, some devices, mostly linguistic or symbolic in nature, that are used in science do so resemble money. Even though scientific languages are different from vernaculars, its choice is only a matter of pragmatics, namely, the effectiveness of invoking the right fundamental representations. However, other devices, mostly not symbolic, are not necessarily of this sort. Whether the point-mass model for our solar system or a ball-stick model for DNA molecules qualifies as a legitimate model, the choice is emphatically not a matter of pragmatics. Notice, the question is not whether the models should be made of steel or plastics; no, that would indeed be a question of pragmatics and convention. The question is what makes the structures with such-and-such components and relations legitimate candidates for representing their targets.

So, first, there may always be a pragmatic aspect about any candidates for modeling, and yet it cannot be the only aspect; and second, it is in most cases not even the relevant or significant aspect; it would never even occur to the model-builders that this is the constitution question that may trouble them in their more philosophical moments. To the relevant question that does concern whether the structures of the models are at least in the ballpark of the target systems, only non-pragmatic considerations figure prominently the answer.

Now, what is the difference between epistemic virtue and the pragmatic one? One may argue that even if I am right above that the epistemic virtue should figure in an answer to our question about the nature of scientific representation (the world around us is represented to us so that we may, among other things, come to *know* it), it still doesn't touch deflationism because epistemic virtue should be understood as part of the pragmatics. This line of thought may go as follows. Yes, one of the chief purposes of using artifacts to model systems in nature is to eventually obtain knowledge about the latter. This is indeed the business of epistemology, the use of models as representational devices is precisely to serve the epistemic purposes, which means what is and what is not a legitimate device to use is a matter of what does or does not serve the purpose of obtaining knowledge through it.

Reasonable as it is, it still misses an important difference. The upshot about being a matter of pragmatics or convention is that “right or wrong” or “truth or falsity” cannot be its concern; it can only be a question of “good or bad” or “useful or obsolete.” A steel model of human DNA is a good model for the epistemic purpose because it is durable and easily recognizable, etc, while a model of it made out of ricotta cheese is bad because . . . obvious reasons. Similarly, a model described and published in English is better for epistemic purposes than one described and published in, say, Chinese for obvious reasons. What is said in the previous section applies to this point; however, we know that this is not the most relevant or significant question about model-building.

When it is a legitimate concern in considering model-building in general whether the results reveal to us to some extent what their targets *are like*, we know that it is no longer only a matter of pragmatics. To further argue for this line of thought, let us turn to the second point.

Why are such considerations not pragmatic? What sort of considerations are they? I want to argue that they are considerations about knowledge acquisition and concerned with *epistemic values*. Let us begin by reflecting on what the most basic model-building process looks like. How does (or might) a human represent a cat for the first time?[7] She notices the presence of something X that she later will recognize, or others around her already recognize, as a cat. She obtains perceptual experience of X, and then either by bold conjecture or by painstaking inductive generalization, forms a complex belief state in her head that she, if reasonably equipped with the skill of mind-eye-hand coordination, may be able to draw or sculpt into a artifact – a model cat – that at the most primitive level recreates a perceptual experience that simulates her earlier experience of X.

Lest I am misunderstood, let me mention quickly and briefly that there is no need to think as a consequence of the above that her experience of X must ‘resemble’ X itself, whatever that may mean, nor is it necessary to conclude that whatever drawings or sculpture she may produce to show what she believes how the cat looks like must ‘resemble’ that particular complex belief state, whatever that may mean. (Perhaps, we can say that the bottom line is that the impression she gets by looking at her drawing or sculpture must be similar to a great extent to the impression she gets from looking at that cat, which is obviously not true if she is looking at the word ‘cat.’) Whatever their nature, these two relationships are likely to be less than straightforward, and discoveries in cognitive science and neuropsychology are likely to continuously revise whatever philosophical accounts we have of them.

So in sum, I demand in this epistemic value no simple-minded stipulations of any relationship of resemblance or similarity or isomorphism between the models and their targets; but I do insist that some sort of natural (or naturalistic) relationship – of whose nature only future empirical research may provide definite answers – holds between the external devices and the belief states they correspond and the belief states and the target systems that the agent is trying to represent, such that when the relationship is not there, the representation must be deemed illegitimate. The

[7] The question is obviously meant in a conceptual way; it is not intended to be a historical question.

holding or not holding of such relationships as being determined by natural and social conditions, not by in principle conventional means, is the content of the epistemic virtues I have been trying to argue for important aspect of the constitution question regarding the business of model-building in science.

Here I find Locke quite prescient in saying the following.

> To discover the nature of our ideas the better, and to discourse of them intelligibly, it will be convenient to distinguish them, as they are ideas or perceptions in our minds, and as they are modifications of matter in the bodies that cause such perceptions in us; that so we may not think (as perhaps usually is done) that they are exactly the images and resemblances of something inherent in the subject; *most of those sensation being in the mind no more the likeness of something existing without us than the names that stand for them are the likeness of our ideas, which yet upon hearing they are apt to excite in us.* (Locke 1796, p. 111, my italics)

Locke is no doubt right about how resemblance or similarity between the representation and the represented cannot be taken seriously as a legitimate constraint on mental representation, he is, however, not to be taken as meaning to say that the ideas or perceptions in our head are no less arbitrary and conventional as words or names we used to stand for them. On the contrary, what mental devices are used by cognitive agents like human beings to represent the world around them must be determined naturally, and moreover the external devices we use to show what the represented are like must also not be arbitrary or conventional. The reason for this latter point is what I have argued just now.

Therefore, the above with the caveat, though naïvely put, must be the origin of scientific representation that takes modeling to be its core task. Whether the scientists are constructing models of the observable or the unobservable systems, the models may or may not be physically realizable, one of the most important conditions of adequacy must be that the model serves as the (imagined) object or cause that by clearly understood ways reproduces the (possible) perceptual experience that resembles that *supposed* original experience.[8] So, resemblance does play a crucial role in scientific representation, but where it plays such a role is usually misunderstood in the literature. It is not that the model should resemble or be similar to the represented system (whatever that means), but rather that what the model are conjectured to produce, by well understood possible causal processes, must be similar or resemble the supposed actual experience about the target system. Here lies the epistemic values of scientific models and they are central to the selection of such models.

[8] For observable systems, this claim can be understood straightforwardly, as in the case of representing a cat one sees for the first time; but for unobservable systems, systems such as atoms and extinct creatures long ago, the resemblance relationship can only be understood as holding with the hypothetical "original" experience, something we imagine by reason of analogy that we could experience if we were put into the supposed circumstances. In the case of the unobservables, the assumption of such a resemblance relationship is in fact more important than with the observables because the models, if deemed correct, would be the only things that could tell us what the target systems may "look like." What else could tell us what the hydrogen atoms look like apart from their latest quantum mechanical model?

A lot more need to be qualified about the above before we can reach a conception of scientific models and/or representation that is applicable to a wide range of practices in science, and I want to say that many of the efforts in the existing literature on the subject, which deflationism has relegated to the bin of 'only dealing with further pragmatic questions of modeling,' have already made great stride towards a comprehensive and sophisticated understanding, which this small paper has no space to enumerate (see also e.g., Hesse 1966; Hughes 1997; Suáres 2003).

One objection to my picture of scientific representation is that to hold that I may have to exclude conventional elements and matters of pragmatics. If what scientific models are intended is to tell us what their targets are like in a substantive epistemic sense, and the legitimacy of one candidate over another for modelhood is determined by such epistemic virtues, then what's the merit of Callender and Cohen's work that I endorsed earlier? To respond, let us first notice something general about all representations, linguistic, artistic or scientific. While all styles of representation have pragmatic aspects to them, each however are primarily aimed at promoting a different set of values while representing the world to us. Artistic representations are for aesthetic virtues and scientific ones for epistemic virtues, while linguistic ones are purely for pragmatic virtues. This said, it is hard to miss the fact that pragmatic values permeate all styles of representations. We may take this as a result of the necessary use of 'language' or 'symbol' in any style of representation: artists need a 'language of art' and scientists a 'language of science.' Therefore, it is not surprising to see how pragmatic values figure in an enterprise that is primarily epistemological (or aesthetic). For instance, our cat representer mentioned above needs to choose a means of representation in order to show others what a cat is like to her. Between the choices of making a simple drawing and reconstituting a cat with flesh and blood, guess which one she is likely to choose under normal circumstances? And the choice is certainly the result of pragmatic considerations. In general, scientists and their communities make pragmatic considerations in choosing external representation vehicles all the time, and as I mentioned above, even with well-worn examples of a point-mass model for solar system or a ball-stick model for DNA molecules, pragmatic concerns are involved between constructing a steel or a plastic structure; but these are obviously not the primary or even the relevant concerns of model-building in the given contexts. Epistemic virtues would obviously be the main concern.

Given how we represent what's around us in our mind, scientifically or otherwise, we must use external devices to show and communicate our representations to others.[9] We can do this for purely pragmatic purposes, just having to make sure others know what we have in mind, or we may do it to show what we think the targets of our representation are like. The making of the latter has to be

[9] This would not be true if non-representationism as mentioned above is adopted. For lack of space, I have to omit any discussion of this point. The conclusion is the same, namely, it is epistemic virtues, not pragmatic ones, which primarily govern the choices of scientific representation.

constrained by the practicality of model construction, but the purpose is primarily epistemic. I therefore suggest that what we usually mean by scientific models are such representational devices; they are not anything that we can use by conventional stipulations/agreements.

References

Burgess, A., & Burgess, J. P. (2011). *Truth*. Princeton: Princeton University Press.

Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria, 55*, 7–25.

Freeman, W. J. (1991). The physiology of perception. *Scientific American, 264*(2), 78–85.

Grice, P. (1989). *Studies in the way of words*. Cambridge: Harvard University Press.

Hesse, M. G. (1966). *Models and analogies in science*. Notre Dame: University of Notre Dame Press.

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science, 64*, S325–S336.

Locke, J. (1796). *An essay concerning human understanding* (20th ed., Vol. 1). London: Thomas Longman.

Morgan, M. S., & Morrison, M. (Eds.). (1999). *Models as mediators*. Cambridge: Cambridge University Press.

Siegel, S. (2011). *The contents of visual experience*. Oxford: Oxford University Press.

Suáres, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science, 17*, 226–244.

Teller, P. (2001). Twilight of the perfect model model. *Erkenntnis, 55*, 393–415.

van Fraassen, B. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Clarendon.

Idealization in Physics Modeling

Demetris Portides

Abstract I argue for understanding idealization in physics modeling as the conceptual act of exercising control over the variability of aspects of factors of target physical systems. By thinking along this line I use the example of the vibrating string model to identify the two kinds of idealizations that have been discerned by most philosophers to be pervasive in scientific modeling, which I label isolation and stabilization, and argue that they are the result of the same kind of thought process. Furthermore, I argue that isolation and stabilization do not exhaust the idealizations that we encounter in scientific modeling. What we need in order to make sense of much of the modern modeling practices, such as quantum mechanical modeling, is the idea of idealization as decomposition.

1 Introduction

Idealization is admittedly pervasive in science and more particularly in scientific modeling. It gives rise to several questions of philosophical interest, among them: What are its functions? What is the nature of its conceptual products? What impact does the presence of idealized ingredients have on the truth-value of scientific hypotheses? What is the nature and character of idealizing processes? In this paper I focus on the latter question. Of course, the answer we give, to what the character of idealization is, is linked to how the various kinds of idealized conceptual products function in scientific models. This makes the attempt to discern the character of idealization processes abstracted from other issues related to idealization all the more difficult.

D. Portides (✉)
Department of Classics and Philosophy, University of Cyprus, P.O. Box 20357,
Nicosia 1678, Cyprus
e-mail: portides@ucy.ac.cy

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_9,

A further difficulty arises from the observation that the products of idealization are heterogeneous. This has led some authors to defend the view that there are two parallel thought processes at work in scientific modeling. For example (Cartwright 1989) and (Suppe 1989) claim that the process of abstraction is a clearly distinct thought process from that of idealization. According to both authors, and others who roughly align themselves with this view, abstraction involves omission of features from the model description whereas idealization involves the distortion of the features of the target system present in the model description.

In my attempt to focus on the thought process that underlies idealization in scientific modeling I defend a conception of idealization that treats 'abstraction' (i.e. omission) and 'idealization' (i.e. distortion) as two particular modes of the same thought process. The thought process involved in modeling is what could be called *conceptual control of variability*. That is to say, the view of idealization I advocate conceives idealization as the thought process by which scientists conceptually exercise 'control' over the variability of factors that involved in the behavior of a target system. Of course, variable control needs clarification when not used in reference to experimental science, thus further analysis is needed in order to shed light on its connection to the notion of idealization. Nevertheless, in this paper my attention is restricted to mathematical models and since it is imperative that we can assign numerical values to the parameters (whether representing factors of influence or properties of factors) involved in such models, it is not vague to speak of the variability of these parameters. In this paper I attempt to analyze and explicate the three general ways by which modelers conceptually control variables, thus it is an attempt to highlight the three kinds of idealization present in scientific models.

In Sect. 2 I elaborate on a well known example from elementary mechanics that I use in order to emphasize that the sense of both abstraction and idealization as employed by Cartwright, Suppe and other philosophers are no more than specific ways by which variable factors of the target system are controlled. In Sect. 3 I focus on a kind of idealization (a third way by which variable parameters of the target are controlled) that has not received much attention in the literature, but which in my view is also pervasive in modeling. I label it decomposition. In order to shed light on the notion of decomposition I try to show how it is involved in the construction of a model from nuclear physics.

2 Isolation and Stabilization

To model a flexible stretched string (much like the ones found in musical instruments) by the use of Newtonian mechanics physicists invite us to imagine a string of length L and mass per unit length μ, the ends of which are attached to perfectly rigid supports holding the string under tension. Of course, such properties as having uniform mass distribution or having perfect rigidity of the supports are something that could only approximately be achieved in an actual apparatus. The latter would involve some impurities in the material of the string that would influence the uniformity of the mass distribution, and the actual supports of the apparatus would be somewhat flexible. Such simplifications, however, enable the physicist to conceptually screen off causal influences from factors that complicate the phenomenon thus focusing on the particular factor that the model is meant to explore. The idealizations involved in the modeling process do not however end here; Newtonian mechanics require an equation of motion for the target system. Newton's 2nd law dictates that the acceleration of a small segment of the string is equal to the net force on the segment. If we consider an arbitrary deformation of the string and focus on an arbitrary segment between x and $x+\Delta x$, the displacement from the horizontal axis at time t is a function of two variables $y(x, t)$. The mass of the segment is $\mu\Delta x$, thus the mass times the acceleration can be written as:

$$\mu\Delta x\frac{\partial^2 y}{\partial t^2} \tag{1}$$

This expresses the y-component of the net force on the segment. Now, the transverse force exerted on the segment by the neighbouring parts of string depends on the slope of the string at the two ends of the segment. Let θ_1 and θ_2 be the angles of the string with the horizontal axis at the left and right ends of the segment. Then

$$\tan\theta_1 = \left(\frac{\partial y}{\partial x}\right)_x \textit{ and } \tan\theta_2 = \left(\frac{\partial y}{\partial x}\right)_{x+\Delta x} \tag{2}$$

The y-component of the force on the segment due to the adjacent string parts is $T\sin\theta_2 - T\sin\theta_1$, where T is the tension in the string. For small angles the difference between $\sin\theta$ and $\tan\theta$ can be neglected. If other forces acting on the segment are also neglected (e.g. the weight of the segment or air resistance on the segment's motion), then Newton's 2nd law can be written as

$$\mu\Delta x\frac{\partial^2 y}{\partial t^2} = T\left(\tan\theta_2 - \tan\theta_1\right) = T\left[\left(\frac{\partial y}{\partial x}\right)_{x+\Delta x} - \left(\frac{\partial y}{\partial x}\right)_x\right] \tag{3}$$

If in the above equation we divide through by $\mu\Delta x$, in the limit as $\Delta x \rightarrow 0$, it becomes:

$$\frac{\partial^2 y}{\partial t^2} = \frac{T}{\mu}\frac{\partial^2 y}{\partial x^2} \tag{4}$$

Equation 4 is a form of the scalar wave equation with well-known analytic solutions. The above modelling process is much indicative of how the general principles of theory are applied in order to model the target system. To propose the wave equation as a candidate for representing the vibrating string, in addition to the initial assumptions, it is clear that physicists leave out much information about the actual experimental apparatus. Namely, the weight of the string segment, that the segment experiences friction, that the motion of the string is impeded (e.g. due to the medium), that the tension in the string is a variable parameter and so on. Introducing these idealizations, in the process of setting up an equation of motion is necessary to arrive at the desirable goal, which is none other than setting up a tractable equation of motion for the representation of the target system.[1]

Moreover, if just one of these idealizing assumptions is removed the consequence is that we are led to an intractable equation of motion. Consider, as an example, the case where the tension is not assumed to be a constant parameter (this would physically imply that the mass distribution is also a variable parameter), then dividing Eq. (3) above with $\mu\Delta x$ and taking the limit as $\Delta x \rightarrow 0$ does not result in the wave equation, since the tension and mass distribution would be functions of the variable x and thus differentiable with respect to x. In such a case the result would be the following equation of motion, for which analytic solutions do not exist for any choice of functions of $T(x)$ and $\mu(x)$:

$$\frac{\partial}{\partial x}\left(\left(\frac{T}{\mu}\right)\frac{\partial y}{\partial x}\right)=\frac{\partial^2 y}{\partial t^2}\Rightarrow\frac{\partial}{\partial x}\left(\frac{T}{\mu}\right)\left(\frac{\partial y}{\partial x}\right)+\left(\frac{T}{\mu}\right)\frac{\partial}{\partial x}\left(\frac{\partial y}{\partial x}\right)=\frac{\partial^2 y}{\partial t^2} \tag{5}$$

This point is indicative of the fact that if physicists were to introduce all the factors that are known to be responsible for the observed behavior of the target system into the model, it would result in an intractable equation of motion. If such an intractable equation could be solved by an appropriate method of approximation then the model could become a predictive tool but at the same time it would be an ineffective epistemic device, since it could not be used to gain much physical insight into the workings of its target system. Hence physicists work with the wave equation as a representation of the vibrating string, because despite the idealizations involved in its construction it can be used to gain epistemic access to its target. This goes to show why idealization as a form of simplification is not just scientifically useful, but also necessary for applying the general principles of theory. But the question I earlier promised to focus on in this paper concerns the character of such idealizations. In other words, is there a way to understand such a variety of idealizing assumptions as being the product of one general kind of conceptual act? I think, yes.

Let me begin by summarizing the main idealizations involved in the different stages of constructing the model for the vibrating string: We assume uniform mass distribution for the string, perfect rigidity of the supports, no weight for the string segment, no frictional forces acting on the segment, no impedance on the motion of the string by the medium inside which it vibrates, and constant tension in the string. These idealizing assumptions have traditionally been divided into two different categories. In Mcmullin's (1985) account all of the above are labeled formal idealizations which involve either simplified descriptions or omitted features.

[1] I take the introduction of such idealizing assumptions in the modelling process to be one of the aspects of what (Cartwright 1983) has dubbed ‘theory entry’. Notice that in order to set up the equation of motion of the system approximation assumptions are also involved, e.g. that θ has a magnitude for which $\tan\theta = \sin\theta$.

The assumptions of uniform mass distribution and of the tension being a constant parameter are cases of what Mcmullin would call formal idealizations that involve simplifications. The assumptions of no weight for the string segment, of perfect rigidity of the supports, that the segment experiences no friction, and that the motion of the string is not impeded by the medium inside which it vibrates are all cases of what Mcmullin would call formal idealizations that involve omission of features.

In all cases, however, Mcmullin construes idealization as a simplification or omission of features present in an actual situation that leads to simplified concepts or simplified descriptions of a situation. Plain omission of features is a straight forward case for Mcmullin, e.g. omitting the presence of a medium inside which the motion of the string in the above example takes place, which amounts to omitting the inclusion of all the effects of such a medium from the model description. What he calls simplification is the kind of idealization where features that are retained in the model description are represented in the model equation in a more simplified way than the way they are perceived in the target system, e.g. the constancy of the tension. Mcmullin blends simplification in the latter sense and omission into his generic notion of idealization. In fact, many authors blend the two notions. For instance, (Nowak 1980) blends the two into his notion of 'idealization'. Similarly, (Morrison 1997) blends them into her notion of 'computational idealization', and (French and Ladyman 1998) also blend them into their notion of 'idealization'.

Other authors (e.g. Cartwright 1989; Suppe 1989), as I mentioned earlier, distinguish these two categories of idealization on the grounds that they are the result of two distinct thought processes, labeling the kinds of assumptions involved in the 'omission' category *abstractions* and those involved in the other *idealizations*. Cartwright distinguishes the two on the grounds that idealizations involve the rearrangement of the mathematically-inconvenient features kept in the model, whereas abstraction simply involves subtraction of features. I have no dispute with the distinction per se. The point I make is that we need not make sense of these different features of models as if they result from distinct thought processes. In fact, although admittedly there are differences between the two cases, these differences do not warrant the conclusion that the two kinds are the result of distinct thought processes. It is possible to have a unifying view of all kinds of idealizations through a particular reconstruction of the conceptual act behind idealizing assumptions. This is another way of saying that it is possible to think of what Mcmullin calls 'simplification' and 'omission' as results of one and the same kind of conceptual act. To do this, I claim, one must rely on what I earlier called control of variability, which I explain below.

In the model of the vibrating string above, as mentioned we assume uniform mass distribution for the string, constant tension in the string, perfect rigidity of the supports, no weight for the string segment, no frictional forces acting on the segment, and no impedance on the motion of the string by the medium inside which it vibrates. These assumptions can easily be reconstructed as conceptually exercising control, of one sort or another, over the variability of the various characteristics of the target system. Assuming uniform mass distribution is equivalent to conceptually rendering the variable 'mass distribution' a constant parameter. Similarly assuming that the tension is constant is equivalent to conceptually rendering the variable 'tension' a constant parameter. Furthermore, assuming perfect rigidity of the supports is equivalent to conceptually setting the variable of 'flexibility of solid joints' to the value of zero. Assuming no weight for the string segment is the same as conceptually setting the variable of 'weight of segment' to the value of zero. Assuming that no frictional forces act on the segment is equivalent to conceptually setting the value of the variable 'friction between segments' to the value of zero. Assuming no impedance of the motion by the medium is equivalent to conceptually setting the variable 'density of medium' to the value of zero. And so on.

All the above are particular ways by which scientists conceptually control the variability of parameters in models. The second set of assumptions that has been identified with abstraction in the sense of omission is no more than conceptually setting the value of a variable parameter to zero which ensures that the particular parameter will not appear in the model description. The first set of assumptions that has been identified with idealization in the sense of distortion is a many-sided form of variable control. The model inevitably must speak about something. Hence a decision must be made about which things to retain and from which unwanted things to conceptually isolate them from. In addition, the mathematical calculus of the theory imposes a further difficulty: in order for the calculus to lead to useful results the things it will describe must have particular characteristics. Otherwise either the calculus is useless in setting up an equation for the system or it leads to intractable equations. Therefore, the characteristics of those things retained in the model description must be structured according to the demands of the calculus. But the characteristics are variable parameters and conceptually modifying them in order to give them the required structure can only be done by controlling their variability. Sometimes this control is exercised by setting a particular variable quantity to a constant value, e.g. the case of the mass distribution and the tension above. Other times the control is achieved by choosing a particular functional expression by which to characterize a variable quantity, such that, the application of the calculus is facilitated, e.g. if we were to chose a particular function by which to express the tension.

It is clear that all idealizing assumptions can be translated into some form or other of control over the variability of the constitutive factors of the target. It is also clear that there are two rather general categories of variability control. The first could be translated into the idea that the (variable) contribution of a particular factor to the behavior of the target is not important to the epistemic function of the model. The second could be translated into the idea that the *significance* of the variability of a particular factor to the behavior of the target is not important to the epistemic function of the model. I choose to call the first kind of idealization *isolation*, since it amounts to conceptually screening off the model content from factors that either exert influence on the observed behavior of the target or not.

And I will call the second kind of idealization *stabilization*, since it amounts to conceptually rendering the variable influence of the parameters retained in the model on the observed behavior of the target uniform or stable.[2]

3 Decomposition

Although the example of the vibrating string is useful for highlighting the above two kinds of idealization it does not help to discern a third kind that more often than not occurs in scientific modeling. Quantum mechanical modeling methodology, for example, cannot be fully understood by relying only on the above two kinds of idealization because it is faced with a further problem: the factors and their interactions responsible for the target system's behavior are convoluted and we are often not able to separate them and observe their individual effects experimentally. This additional problem leads to a third kind of idealization involved in several areas of scientific modeling, such as quantum mechanics and evolutionary theory, which has not received much attention in the literature and which I will label *decomposition*.

Decomposition consists in setting apart various clusters of influencing factors, resulting in a description that involves distinct clusters of factors thought to be acting in tandem to produce the particular behavior of the target system. Generally speaking, idealization by decomposition can be sub-divided into four types. The first two are the obvious cases which we come across in almost all scientific domains, i.e. decomposition of the effects of a particular cause and decomposition of the causes of a particular effect. For example, we know that if we were to incorporate into the vibrating string equation the effects of the medium we would have to account for how the medium impedes the string, for how the string is affected by buoyancy and for how the string is forced to partially rotate along the axis of vibration. These are no more than some decomposed effects of the same cause – the medium. Despite the knowledge that whenever vibrations experimentally occur in a medium these effects occur together, we decompose them in our model descriptions as if one of them could occur in isolation from the others. Of course decomposing the effects of a common cause or the causes of a particular effect is so compatible with our intuitions and with much of our experimentation, e.g. we think of sunlight as being composed of seven discrete colors because of experimentation, that we tend to overlook that often we treat things as if they are decomposed even though no experimental evidence corroborates that. In such cases it is, I think, best to think of decomposition as a form of idealization.

When we focus on the last two types of decomposition that we often come across in scientific modeling, which are the ones I am primarily concerned with,

[2] I borrow the term 'isolation' from (Mäki 2011), in which it is used to express a rather similar idea, and the term 'stabilization' from (Zielinska 1990), in which it is used to express a related idea.

it becomes rather clear that decompositions are kinds of idealizations. The third type of decomposition leads to a number of distinct model-representations each responsible for a particular behavior of the target system (i.e. construction of multiple-special-case-models). This type is common to both physics and biology. In evolutionary biology, for example, for a period models were constructed to treat the mechanism of natural selection as if it acts on its own and independent of other processes. Similarly, other models were constructed to treat the mechanism of genetic drift also as if it acts on its own. For a part of the history of evolutionary theory such special case models were constructed to treat the processes of selection, drift, mutation and migration as being separate and independent from each other. These are clear cases of models that are underpinned on decomposition.

Having decomposed the evolutionary process into selection, drift, mutation and migration more recently biologists devised diffusion theory, which is the tool by which they attempt to bring together into a single model the interactions of these separately-treated processes and thus study their joint effects. However, in evolutionary models based on diffusion theory the processes of selection and drift, for example, are still decomposed, despite the fact that their interaction is accounted for. In other words, decomposition of processes also occurs within the same model, this is the fourth type. In such cases, a description of the target system is constructed in which clusters of factors, that are all constitutive parts of the same representation, are thought to act independently in order to produce the observable effects jointly. This is what occurs in diffusion models in evolutionary biology, and this is what most often occurs in quantum mechanical models. Another way to put this point is this: a general characteristic of models is that they are used to explain the observed behavior of their targets by assuming that the behavior is the joint product of separate but interacting processes. Of course this may actually be the case in many instances, but my point is that very often we do not know if it is the case and other times we have no means of knowing, thus what underlies such model-descriptions is decomposition, i.e. a kind of idealization that presents a simplified picture of the target by explaining the observed behavior as if it is the result of parallel and independent processes. I will sketch the case of the *liquid drop model* (also known as the collective model) of nuclear structure to clarify this kind of idealization.[3]

The liquid drop model of the nucleus assumes that the nuclear properties are produced by the collective behavior of nucleons. In other words, it omits any independent nucleon contribution to the observed nuclear properties. This of course is indicative of the fact that in the construction of the liquid drop many idealizations of the kinds I have called isolations and stabilizations are involved. Let me, in what follows, abstract from those in order to focus on the aspect of decomposition. The Hamiltonian operator of the liquid drop consists of four distinct terms. The first

[3] Detailed expositions of the liquid drop model of nuclear structure can be found in (Eisenberg and Walter 1970) and in (Moszkowski 1957). In (Portides 2006) I present an analysis of the importance of the liquid drop model in the evolutionary history of models of the nuclear structure.

term is due to a rotation collective mode of motion, the second due to a vibration mode, the third due to a combined rotation-vibration mode and the third due to giant resonance:

$$H_{COL} = H_{ROT} + H_{VIB} + H_{ROT-VIB} + H_{GR}$$

Each of the above four operators is a cluster of several factors that are expressed by making several different idealizing assumptions of the kinds I have called isolation and stabilization. Since I am abstracting away from those and focusing on decomposition it is unnecessary to look at the detailed expression of each of these operators. The above Hamiltonian operator expresses a particular conception of the potential energy operator the details of which I shall ignore. The point is that the nuclear potential is presented in the above Hamiltonian as if it is due to a strongly coupled collection of nucleons which rotates, vibrates, rotates-vibrates and exhibits a mode of motion known as giant resonance. The latter is a particular kind of phenomenon that we come across in nuclear physics, the details of which play no role to my argument.

Now, the liquid drop model is successful in explaining, and to a first approximation gives good predictions for, the phenomena of nuclear fission and the electric quadrupole moments of nuclei. But if we think, for the sake of argument, in classical terms it is not possible for us to experimentally determine whether e.g. nuclear fission is the result of distinct modes of motion of rotation, vibration and so on. The claim is that the decomposition, within the model, of the modes of motion of the strongly coupled collection of nucleons is conceptual. Actual nuclei have some complex mode of motion which we conceptualize as being produced by the above four distinct motions integrated together. This, however, is the result of a conceptual act that does not necessarily correspond to the realities of actual nuclei.[4]

Idealization as decomposition is the conceptual act of clustering parameters of the target and setting apart the clusters with the assumption that each cluster describes a process (or mechanism) independent from the others, and that these processes act in tandem to produce the observed results. In other words, the underlying assumption is that the observed result is no more than the joint effect of the clusters of factors of the model. Generally, we can think of idealization as decomposition as the conceptual act of setting apart descriptions of different processes that are assumed to act independently from each other. Can decomposition be reconstructed and conceptualized as a form of variable control? I think yes.

[4] Notice that I ignore the fact that the liquid drop is what physicists would call a semi-classical model, because I am not concerned with whether the model presents a realistic representation of the nucleus or whether the model establishes a good or bad approximate link between quantum theory and experimental results. Rather, I am concerned with the presence of decomposition in the model's Hamiltonian operator. And since decomposition is present in even more elaborate and realistic models of the nucleus, such as the unified model of nuclear structure, and also in other quantum mechanical models such as the BCS model of superconductivity, the choice of using a semi-classical model to shed light on idealization as decomposition is just the outcome of the quest for clarity of explanation.

The variability in this case is an aspect of the interconnection between factors. When we include two or more parameters into the same cluster then we account for the interconnection of the two, i.e. how the two interact and how they may be related. That is to say, how each of the constituent parts of each of the individual components of the liquid drop Hamiltonian, i.e. H_{ROT}, H_{VIB}, $H_{ROT\text{-}VIB}$, H_{GR}, interacts with the rest of the constituents in the same cluster is accounted for. However, when we include two or more clusters into the model we do that by abstracting away from their interconnection, i.e. we control the variability of their interconnection by setting it to zero. This, of course, translates into the idea that the significance of the interaction between different processes is considered not to be important for the epistemic function of the model. Of course, modelers know too well that even if such separate clusters in the model are realistic descriptions of aspects of the target, that in order to obtain a more accurate representation they must account for the interactions between them. In quantum physics normally this is done by introducing a separate interaction term in the Hamiltonian operator often by the use of perturbation theory or by other means. In biology, as I mentioned earlier, this is done by the use of diffusion theory. Nevertheless, I interpret the introduction of interaction terms in models that have separate clusters of components as corroborating my conclusion: that idealization by decomposition into processes (that are assumed to act independently) has to be overcome in one way or another if a model of such sort is to say something more realistic about its target.

4 Conclusion

I have argued for understanding idealization in physics modeling as the conceptual act of exercising control over the variability of aspects of factors of target physical systems. Along this line of thinking I have identified two kinds of such idealizations (i.e. isolation and stabilization) that have been discerned by most philosophers to be ubiquitous in science, and argued that they are the result of the same kind of thought process. Furthermore, I have argued that isolation and stabilization do not exhaust the idealizations that we encounter in scientific modeling. What we need in order to make sense of much of the modern modeling practices is the idea of idealization as decomposition.

References

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon Press.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Clarendon Press.

Eisenberg, J. M., & Walter, G. (1970). *Nuclear theory: Nuclear models* (Vol. 1). Amsterdam: North-Holland.

French, S., & Ladyman, J. (1998). Semantic perspective on idealisation in quantum mechanics. In N. Shanks (Ed.), *Idealisation IX: Idealisation in contemporary physics, Poznan studies* (Vol. 63, pp. 51–73). Amsterdam: Rodopi.

Mcmullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science, 16*, 247–273.

Mäki, U. (2011). Models and the locus of their truth. *Synthese, 180*, 47–63.

Morrison, M. (1997). Models, pragmatics and heuristics. *Dialektik, 1*, 13–26.

Moszkowski, S. A. (1957). Models of nuclear structure. In S. Flügge (Ed.), *Encyclopedia of physics: Structure of atomic nuclei* (Vol. 39, pp. 411–550). Berlin: Springer Verlag.

Nowak, L. (1980). *The structure of idealization*. Dordrecht: Reidel.

Portides, D. (2006). The evolutionary history of models as representational agents. In L. Magnani (Ed.), *Model-based reasoning in science and engineering, texts in logic* (Vol. 2, pp. 87–106). London: College Publications.

Suppe, F. (1989). *The semantic conception of theories and scientific realism*. Urbana: University of Illinois Press.

Zielinska, R. (1990). A contribution to the characteristics of abstraction. In J. Brzezinski, F. Cogniglione, T. Kuipers, & L. Nowak (Eds.), *Idealisation II: Forms and applications, Poznan studies* (Vol. 17, pp. 9–22). Amsterdam: Rodopi.

Explanatory Models Versus Predictive Models: Reduced Complexity Modeling in Geomorphology

Alisa Bokulich

Abstract Although predictive power and explanatory insight are both desiderata of scientific models, these features are often in tension with each other and cannot be simultaneously maximized. In such situations, scientists may adopt what I term a 'division of cognitive labor' among models, using different models for the purposes of explanation and prediction, respectively, even for the exact same phenomenon being investigated. Adopting this strategy raises a number of issues, however, which have received inadequate philosophical attention. More specifically, while one implication may be that it is inappropriate to judge explanatory models by the same standards of quantitative accuracy as predictive models, there still needs to be some way of either confirming or rejecting these model explanations. Here I argue that robustness analyses have a central role to play in testing highly idealized explanatory models. I illustrate these points with two examples of explanatory models from the field of geomorphology.

1 Introduction

Prediction and explanation have long been recognized as twin goals of science, and yet a full understanding of the relations – and tensions – between these two goals remains unclear. When it comes to scientific modeling there are two well-known problems with any close marrying of prediction and explanation: First, there are phenomenological models that are highly useful for generating predictions, yet offer no explanatory insight. Hence, predictive power *simpliciter* cannot be taken as a hallmark of a good explanation. Second, as a matter of fact, explanatory

A. Bokulich (✉)
Department of Philosophy, Boston University, Boston, MA 02215, USA
e-mail: abokulic@bu.edu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_10,

power and predictive accuracy seem to be competing virtues in scientific modeling: a gain in explanatory power often requires sacrificing predictive accuracy and vice versa.

In the philosophy of biology, the tradeoffs scientists seem to face between explanatory and predictive models have received some attention, beginning with Richard Levins (1966) article "The Strategy of Model Building in Population Biology" and the various responses to it (e.g., Orzack and Sober 1993; Matthewson and Weisberg 2008). However, these tensions remain an under-explored issue in the philosophy of science and more cases need to be examined.

In what follows I will examine models from a field known as geomorphology, which is concerned with understanding how landforms change over time. There have been a number of interesting debates recently in the geomorphology literature about how to properly model geomorphic systems for the purposes of explanation and prediction. I shall use this work in geomorphology to address the following issues in the philosophy of science. First, I shall argue that there is, what I call, a "division of cognitive labor" among scientific models; that is, even for the same natural phenomenon, different scientific models better serve different modeling goals. Although this division of cognitive labor among models might seem obvious upon reflection, it is rarely explicitly articulated. Making this modeling strategy explicit, however, has important implications, in that it can forestall certain types of criticisms and reveal others. More specifically, it raises the possibility that a model that was designed for the purpose of scientific explanation may fail to make quantitatively accurate predictions (a different cognitive goal). Hence, recognizing that there is division of cognitive labor among models might suggest that criticisms involving the quantitative accuracy of an explanatory model are misguided. This, however, raises the second question that I wish to explore in this paper: How are prima facie explanatory models to be tested, and either accepted or rejected? Here I shall argue that robustness analyses have a central role to play in the testing and validating explanatory models. I shall illustrate these points using two examples of reduced complexity models that are being used to explain phenomena in geomorphology.[1]

2 Reduced Complexity Models: "Reductionism" Versus "Synthesism"

Geomorphology is referred to most broadly as the science of the Earth's surface; it is concerned more specifically with how landscapes change over time, and includes land-water interfaces such as coastal processes. Understanding how and

[1] For the purposes of this paper I will take it as already established that idealized scientific models can be explanatory (see, for example, Bokulich 2011); how and when models should (or should not) be counted as genuinely explanatory is discussed elsewhere (Bokulich 2008, 2012).

why landscapes change over time involves synthesizing information from many different fields, including geology, hydrology, biology, geochemistry, oceanography, climatology, etc. Landscape change is strongly influenced by the relative presence, absence, and kind of vegetation, as well as the behavior of both human and non-human animals.

These sort of complexities have led to a number of interesting debates about the proper way to model geomorphic systems. One of the central debates concerns the appropriate level of scale at which to model geomorphic systems and has given rise to two broad approaches to modeling within geomorphology termed the “explicit numerical reductionist” approach versus the “synthesist” approach.

The traditional approach, which is termed by its opponents “explicit numerical reductionism” – or just “reductionist modeling” – tries to remain as firmly grounded in classical mechanics as possible, invoking laws such as conservation of mass, conservation of momentum, classical gravitation, entropy, etc. Moreover, it seeks to represent in the model as many of the physical processes known to be operating as possible and in as much detail as is computationally feasible. Models that are developed in the reductionist approach are termed “simulation models.”[2] As Brad Murray describes them, “Simulation models are designed to reproduce a natural system as completely as possible; to simulate as wide a range of behaviors, in as much detail, and with as much quantitative accuracy as can be achieved” (Murray 2003, p. 151).

By contrast, the so-called “synthesist” school of modeling in geomorphology, argues that complex phenomena don’t always require complex models. As one of the founders of the synthesist approach, Chris Paola, explains,

> The crux of the new approach to modelling complex, multi-scale systems is that behaviour at a given level in the hierarchy of scales may be dominated by only a few crucial aspects of the dynamics at the next level below. Crudely speaking, it does not make sense to model 100% of the lower-level dynamics if only 1% of it actually contributes to the dynamics at the (higher) level of interest.” (Paola 2001, p. 2)

Rather than appealing to the fundamental laws, this approach tries to represent the effects of the lower level dynamics by a set of simplified rules or equations. These simplifications are not seen as an “unfortunate necessity”, but rather as the proper way to model such complex systems. Indeed synthesists such as Brad Murray argue that understanding how the many small-scale processes give rise to the large scale variables in the phenomenon of interest is a *separate* scientific endeavor from modeling that large scale phenomenon (Murray 2003; Werner 1999).

This division in geomorphology between the “reductionists” and the “synthesists” was arguably precipitated by the introduction of a new breed of models in

[2] The term ‘simulation’ is meant here in the sense of imitating the processes in the real-world system as closely as possible, not whether it is a model run on a computer simulation. This is choice of term is somewhat unfortunate in that both the “simulation” models and the rival “reduced-complexity” models are run as computer simulations.

geomorphology termed “reduced complexity models” (abbreviated RCM). Geomorphologists Nicholas and Quine note that,

> In one sense, the classification of a model as a ‘reduced-complexity’ approach appears unnecessary since, by definition, all models represent simplifications of reality. However, in the context of fluvial geomorphology, such terminology says much about both the central position of classical mechanics within theoretical and numerical modelling, and the role of the individual modeller in defining what constitutes an acceptable representation of the natural environment.” (Nicholas and Quine 2007, p. 319)

An important class of these RCM models are known as “cellular models”, which are distant descendants of cellular automata models (Wolfram 1984). Many geomorphologists point to Murray and Paola’s 1994 cellular model of braided rivers, which was published in *Nature*, as “pioneering” (Nicholas 2010, p. 1) and marking a “paradigm shift” (Coulthard et al. 2007, p. 194) in geomorphic modeling. In the next section I will very briefly outline Murray and Paola’s model of river braiding as an illustration of these reduced complexity models and their controversial successes and failures. I will argue that a proper evaluation of reduced complexity models requires attending to the sort of scientific *uses* to which they are put. More specifically reduced complexity models tend to be most useful for generating scientific explanations, and not for more detailed predictions regarding specific systems.

3 Case #1: The MP Model Explanation of Braided Rivers

Rivers come with several different morphologies: some are relatively straight, others are meandering, and still others are braided. A braided river is one in which there is a number of interwoven channels and bars that are dynamically shifting and rearranging over time, while maintaining a roughly constant channel width (as seen below in Fig. 1).

Murray and Paola succinctly describe the goal and results of their reduced complexity model as follows:

> Many processes are known to operate in a braided river, but it is unclear which of these are essential to explain the observed dynamics. We describe here a simple, deterministic numerical model of water flow over a cohesionless bed that captures the main spatial and temporal features of real braided rivers. The patterns arise from local scour and deposition caused by a nonlinear dependence of bedload sediment flux on water discharge. . . . our results suggest the only factors essential for braiding are bedload sediment transport and laterally unconstrained free-surface flow. (Murray and Paola 1994, p. 54)

Note, in this quotation, that the goal of the Murray-Paola (or MP) model is *explanation* – to explain why, in general, rivers braid – not to predict the specific braided pattern of any given river. The guiding assumption behind this model is that, although there are many processes operating, only a small number of relatively simple mechanisms are needed to produce these complex dynamics.

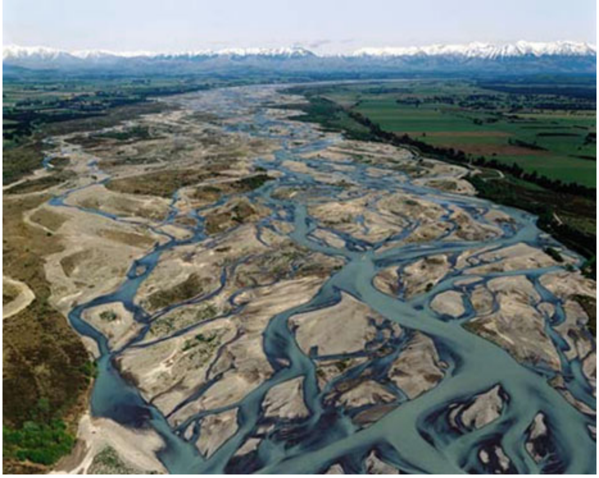


Fig. 1 Example of a braided river: the Waimakariri River in New Zealand

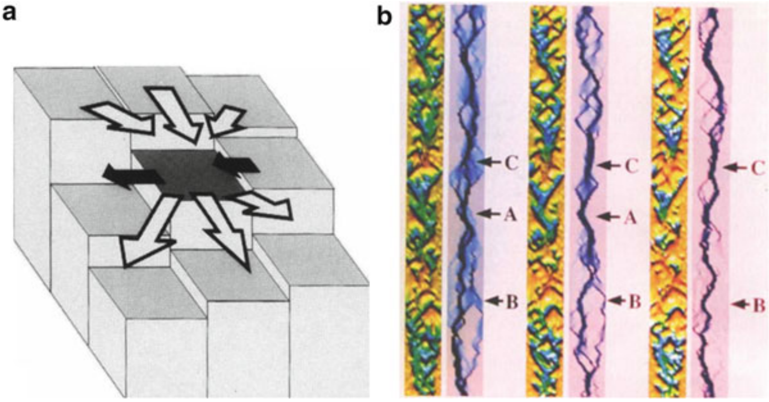


Fig. 2 (**a**) Schematic illustration of the rules in the cellular reduced complexity model of river braiding. The *white arrows* represent water and sediment routing and the *black arrows* show lateral sediment transport (From Murray and Paola 1994, p. 55 reproduced with the permission of the author). (**b**) Three successive times in one run of the model showing 20 × 200 cells with flow from *top* to *bottom*. The l.h.s. of each pair is the topography and the r.h.s. is the discharge (From Murray and Paola 1994, p. 56; reproduced with the permission of the author)

The MP model is essentially a type of "coupled lattice model." In geomorphology, such models represent the landscape – here the river channel – with a grid of cells, and the development of the landscape is determined by the interactions between the cells – here fluxes of water and sediment – using rules that are highly simplified and abstracted representations of the governing physics. Figure 2 shows a schema of the MP model and the patterns this model produces.

On the one hand these reduced complexity models can generate braided rivers with realistic patterns and statistical properties. On the other hand, the simplified rules used to route the water and sediment in the MP model have been called a "gross" simplification of the physics that "neglect most of the physics known to govern fluvial hydraulics (i.e., they do not solve a form of the Navier-Stokes equations)" (Nicholas 2010, p. 1).

It should be noted that there are models of river braiding in geomorphology that do adopt the rival "reductionist" approach, and try to simulate the river in as much accurate detail as is computationally feasible. The so-called DELFT3D model, for example, tries to solve the Navier-Stokes equations in three dimensions and includes many other processes such as the effects of wind and waves on flow and sediment transport. As Murray recounts, "DELFT3D is intended to be as close to a simulation model . . . as is practical, and is probably the best tool available for predicting or simulating fluvial [flow] . . . and bathymetric evolution [i.e., variations in the depth of the river or sea bed]" (Murray 2003, p. 159). However, such models are so complex that they yield very little insight into *why* the patterns emerge as they do.

Murray and Paola defend their highly idealized cellular model by emphasizing that the goal or purpose of their model is not to have a realistic simulation of braided rivers in all their complex detail, but rather to identify the fundamental *mechanisms* that cause a river to braid. Here their model was prima facie successful:

> This simple model showed that feedback between topographical routed flow and nonlinear sediment transport alone presents a plausible explanation for the basic phenomenon of braiding (with lateral transport playing a key secondary role in perpetuating behavior). The model does not include details of flow or sediment-transport processes, such as secondary flow in confluences, and does not resolve distributions of flow and sediment transport on scales very much smaller than a channel width, suggesting that these aspects of the processes are not critical in producing braiding—that they are not a 'fundamental' part of the explanation. (Murray 2003, p. 158)

In other words, the purpose of the reduced complexity model is *explanatory* – to provide an explanation for why, in general, rivers braid by isolating the crucial mechanism.

Moreover, the mechanism for braiding seems to exhibit a sort of *representational robustness* – that is, it is not sensitive to the details of how the sediment-flux "law" or rule is represented: As Murray explains, "[b]raiding is a robust instability in the cellular model, which occurs for any set of rules and parameters that express the non-linear nature of the relationship between flow strength and sediment transport" (Murray and Paola 2003, p. 132). This sort of "insensitivity" of the explanandum phenomenon to the details of the rules or values of the parameters is discovered by performing what geomorphologists call *sensitivity experiments*, which can be thought of as a kind of robustness analysis (e.g., Weisberg 2006a).

In response to the challenge that the availability of these more detailed, physics-based simulation models displaces the need for reduced complexity models such as the MP model, Paola muses,

> Ironically, the debate between synthesism and reductionism has arisen just as the increasing power of relatively cheap computers seems set to make it irrelevant. If we can solve the complete set of primitive equations with a computer, why not just do it? But this debate is

> far more fundamental than mere computing efficiency; it really goes to the heart of what science is about . . . CPU speed may double every 18 months, but the grasp of human intelligence does not. (Paola 2001, p. 5)

There is, however, another issue here that goes beyond the point of limited computing power – human or otherwise – and that is the issue of what makes something a *good* explanation. Arguably a good explanation is one that only includes the essential features needed to account for the phenomenon (for the purpose/context in question).[3] An explanation that includes far more than what is really needed to account for the phenomenon of interest (that is, an explanation that includes the proverbial kitchen sink) is arguably an inferior explanation, quite apart from whether or not the human mind is capable of seeing through those excessive details. The role of reduced complexity models in geomorphology is precisely to isolate just those fundamental mechanisms that are required to produce – and hence explain – a poorly understood phenomenon in nature.

The way the debate has played out between the "reductionists" and "synthesists" suggests that the proper question is not "What is the best way to model braided rivers?", but rather "What is the best way to model braided rivers *for a given purpose*?", where that purpose can be either explanatory insight or predictive power (or indeed something else). As Ron Giere reminds us, "There is no *best* scientific model of anything; there are only models more or less good for different purposes" (Giere 2001, p. 1060). I want to build on this insight and argue that, in geomorphology at least, we can nonetheless identify different *kinds* of models as being better for different *kinds* of goals or purposes. Very roughly, if one's goal is explanation, then reduced complexity models will be more likely to yield explanatory insight than simulation models; whereas if one's goal is quantitative predictions for concrete systems, then simulation models are more likely to be successful. I shall refer to this as the *division of cognitive labor among models*.

Recognizing that there is a division of cognitive labor among models in scientific practice, however, raises its own set of philosophical issues, which have not yet received adequate attention in the literature. For example, a model that was designed for the purpose of generating explanatory insight, may fail to make quantitatively accurate predictions for specific systems (a different cognitive goal). This failure in predictive accuracy need not mean that the basic mechanism hypothesized in the explanatory model is incorrect. Nonetheless, explanatory models need to be tested to determine whether the explanatory mechanism represented in the model is in fact the real mechanism operating in nature. To bring this issue of the testing of explanatory models into focus, let me introduce one more example of a reduced-complexity model explanation in geomorphology.

[3] A full discussion of what distinguishes a good explanation from a poor one is outside the scope of this paper.

4 Case #2: The Model-Explanation of Rip Currents

Another enigmatic physical phenomenon that geomorphologists are using reduced complexity models to explain is rip currents, which are strong isolated offshore-directed flows that appear abruptly at apparently random locations, and last only for tens of minutes before disappearing. These "flash rips" as lifeguards call them are not produced by channels in the sea bed as other, more well understood, rip currents are. Rather, flash rips appear to be hydrodynamical in origin. Rip current velocities can be as fast as 1 m/s and they claim the lives of many beachgoers every year.

In order to explain the origin of these rip currents, Murray and Reydellet begin with the long-known observation that in a very strong rip current, a gap in wave breaking can extend through the surf zone. They furthermore observe that "the waves are not generally larger when they reach the shore than they are in adjacent areas, suggesting that some process other than breaking dissipates wave energy" (Murray and Reydellet 2001, p. 518). They hypothesize that the mechanism for this dissipation could also be part of the mechanism that is responsible for the formation of rip currents on planar (i.e., relatively flat) beaches.

As in our other example, they use a cellular "reduced complexity model" to propose an explanation for how and why such rip currents occur. A schematic diagram of the hypothesized mechanisms leading to rip currents is picture in Fig. 3 below.

The basic idea is that a weak offshore flow begins to decrease wave heights locally, which allows the offshore slope to accelerate the current. Then, as seen in Fig. 3b, the removal of water from the surf zone locally creates alongshore surface slopes that drive alongshore currents feeding the rip current (Murray et al. 2003, p. 270). They show how the interactions between these few simple mechanisms lead to the formation of rip currents in their model simulations. Moreover, they note how

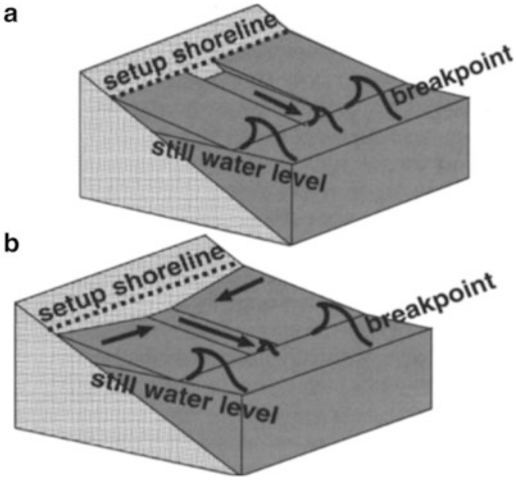


Fig. 3 Schematic illustration of hypothesized wave-current interaction in the reduced complexity model of rip currents. In (**a**) a weak offshore flow decreases wave heights locally, allowing offshore slope to accelerate current. In (**b**) removal of water from surf zone causes alongshore surface slopes that drive alongshore currents feeding the rip (Reproduced from Murray et al. 2003, p. 270 with permission)

a number of unanticipated features of real rip currents emerge in their model, such as the typical narrowness of the rip current and the wide spacing between adjacent rip currents along a beach.

Once again we see these scientists emphasizing a division of cognitive labor among models, and defending the explanatory function of such reduced complexity models; they write,

> In this numerical model some processes have been intentionally omitted, some treated in abstracted ways (e.g., the cross-shore currents) and some represented by simplest first-guess parameterizations (e.g., the newly hypothesized wave/current interaction) in an effort to determine the essential mechanisms causing flash-rip behaviors. The purpose of such a highly simplified model is to find the most concise explanation for a poorly understood phenomenon, not to reproduce the natural system with maximal quantitative accuracy. (Murray et al. 2003, p. 271)

One of the implications of the above quotation is that it is inappropriate to judge such an explanatory model by the quantitative accuracy of its predictions. As Murray has argued elsewhere,

> for a highly simplified model in which many of the processes known to operate in the natural system have been intentionally left out, and others might be represented by simplest-first-guess parameterizations . . . accurate numerical predictions might not be expected. In such a case, the failure of a numerical model to closely match observations would not warrant rejecting the basic hypotheses represented by the model. (Murray 2003, p. 162)

Nonetheless explanatory models need to be tested, and the conditions specified under which the 'basic hypotheses' of the model can be rejected, otherwise such model-based explanations would carry little force.

These geomorphologists are very much aware of what philosophers refer to broadly as the problem of underdetermination. In the geomorphology literature underdetermination is typically discussed under the rubric of "equifinality", following the terminology of the hydrologist Keith Beven (1996). Beven defines equifinality as the problem that, "in modeling . . . good fits to the available data can be obtained with a wide variety of parameter sets that usually are dispersed throughout the parameter space" (Beven 1996, p. 289). More specifically in the present context, there is the worry that these reduced complexity models might just be phenomenological models, that are able to reproduce the right phenomenon, but not for the right reasons. If the mechanisms producing the phenomenon in the model do not correspond to the mechanisms producing the phenomenon in the real world, then such models cannot be counted as being genuinely explanatory.

How are such highly-idealized explanatory models to be tested and validated? Robustness analyses seem to play an important role at two junctures in validating these models. First, as in the case of the model explanations of river braiding, robustness analyses in the form of sensitivity experiments are performed to ensure that the phenomenon in the model is not an artifact of the idealizing assumptions or arbitrary choice of parameters. Here Murray and Reydellet note,

> The results show that strong, narrow, widely spaced rip currents can result robustly from some relatively simple interactions between a small number of processes. Model experiments have shown that this qualitative result does not depend sensitively on model

> parameters, or the details of the treatments of the processes in the model. (Murray and Reydellet 2001, p. 528)

There is, however, a second juncture at which robustness plays a role that is more directly relevant for the issue of confirmation. Murray and Reydellet note that their rip current model can produce *quantitatively* accurate predictions: specifically it produces rip currents with realistic spacings, and typical velocities, widths, and durations that match field data from Doppler-sonar observations. Interestingly, however, they reject this traditional kind of confirmation as carrying much epistemic weight. Instead, they note that this

> quantitative realism relies on the tuning of two poorly constrained parameters, and in a model that represents some of the processes in ways that do not have a track record of use in other models or comparison with independent measurements, being able to tune parameters or adjust the formal way interactions are treated to produce a match might not provide impressive evidence in favor of the model. (Murray et al. 2003, p. 271)

In other words, there are situations in which quantitative accuracy can be bought cheaply and so should not be considered the be-all and end-all of confirmation. Here, instead, they argue that certain kinds of robust *qualitative* predictions carry much more epistemic weight:

> For such a highly simplified model, a different kind of prediction needs to be tested—a prediction that arises robustly from the basic interactions in the model, and does not depend on parameter values or the details of how the interactions are treated in the model. (Murray et al. 2003, p. 271)

They determine two such qualitative tests for this reduced-complexity model explanation of rip currents. The first involves the qualitative prediction that the prevalence of rip currents – which they quantify with a parameter called "rip activity" or RA – decreases with increasing variation in incident wave heights. The second qualitative test involves the prediction that rip currents should be less frequent (and weaker) on beaches that are steeper. In both cases they show that these predictions derive from the fundamental mechanism hypothesized in the model. They then compared these model predictions to field observations of real rip currents on Torrey Pines Beach near San Diego. As Fig. 4 below indicates, results from the video footage showed the same *trend* of decreasing rip activity with increasing wave-height variability displayed in the model.

To test the second prediction regarding beach slope they compared rip activity on Torrey Pines with rip activity on two other beaches in southern California (Carlsbad Beach and San Onofre Beach) with respectively steeper slopes on days when the wave conditions were similar on all three beaches. As Fig. 5 above shows, the field observations once again show the same *trend* as the model predictions.

They conclude that,

> Extensive model experiments indicate that the trends in the model results shown in Figures [4] and [5] do not vary; they result inexorably from the essential interactions and feedbacks in the model. Field observations that did not show the predicted trends could have falsified the model. (Murray et al. 2003, p. 276)

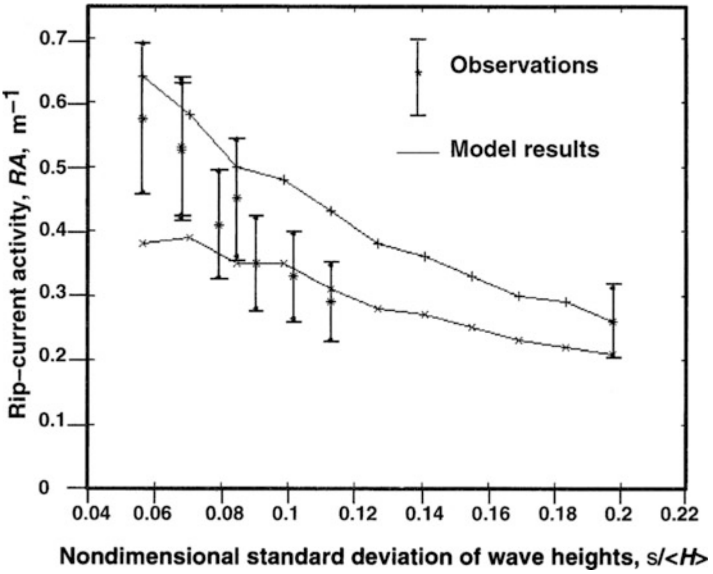


Fig. 4 Comparison of the model prediction that there will be fewer rip currents when there is a greater variation in the heights of incident waves, with observations of actual rip current activity on Torrey Pines Beach, CA

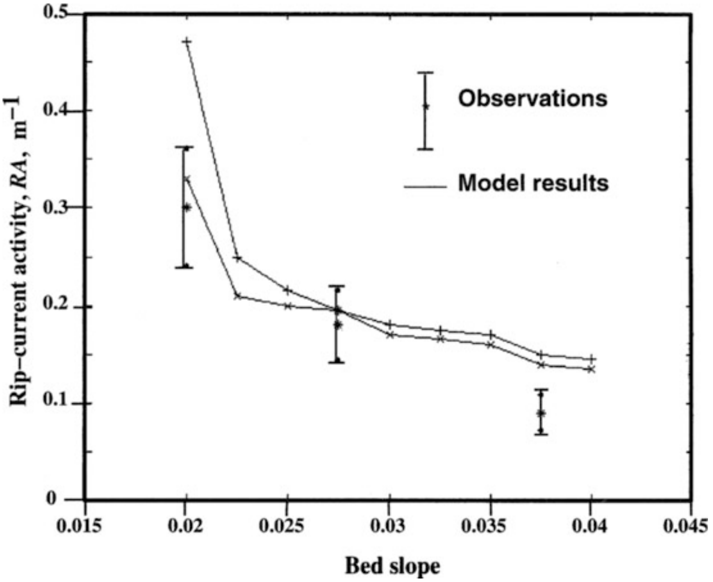


Fig. 5 Comparison of model prediction that rip current activity will decrease when the beach in the surf zone has a steeper slope, with observations from three different beaches of differing slopes in CA

Indeed they note that there are rival models of rip currents that predict trends that are inconsistent with these field observations. They conclude that with these highly idealized explanatory models, *qualitative* predictions involving robust trends can be a more reliable form of model validation.

5 Conclusion: Some Philosophical Lessons from Geomorphology

Several of the themes I have reviewed here in geomorphology relate to similar debates that have occurred in the philosophy of biology. In the context of population biology, for example, Richard Levins (1966) has famously argued that there are tradeoffs between the modeling goals of generality, realism, and precision, and that robustness analyses play an important role in the validation of models. Both of these claims have been challenged by Steven Orzack and Elliott Sober (1993), who question whether there is any necessary conflict between generality, realism, and precision, and who reject robustness analyses as a highly-suspect, non-empirical form of confirmation.

The tradeoff I have described here is not between generality, realism, and precision specifically, but rather between quantitative predictive accuracy and explanatory insight. My aim has not been to argue that there is a *necessary* tradeoff between these scientific goals, but rather to point out that – as a matter of fact – geomorphologists tend to use different kinds of models for achieving different kinds of goals, even when it comes to the same phenomenon in nature. Moreover, we saw intuitively plausible reasons for why reduced complexity models tend to be better for isolating explanatory mechanisms, but worse for generating quantitatively accurate predictions for specific systems, while conversely, the more detailed "reductionist" simulation models tend to be more useful for concrete quantitative predictions, but often are too complex to offer much in the way of explanatory insight.

Despite these important differences, geomorphologists note that "reduced complexity" and "simulation" models mark a difference of degree, not a difference of kind, and that there is a continuous spectrum of models between them. Nonetheless, rather than thinking that there is one, "best" scientific model in this spectrum that is simultaneously optimal for prediction an explanation, geomorphologists instead seem to embrace, what I termed, a "division of cognitive labor" among models, routinely employing different models of the same phenomenon to achieve different epistemic ends.

Turning finally to the issue of robustness, there were two junctures at which geomorphologists were deploying robustness analyses in these examples. The first was in the context of so-called "sensitivity experiments," which were used to determine whether the effect in the model was a robust result of the mechanism(s) hypothesized, or whether it was an artifact, depending sensitively on the values of the parameters. Such sensitivity experiments can be thought of as revealing a kind of *representational robustness*: the phenomenon of interest arises robustly from the mechanism(s) represented in the model, and does not depend on the other idealizations or the particular way in which it is represented in the model.

It is noteworthy that even Orzack and Sober grant the utility of this sort of robustness analysis. They write,

> So far we have considered robustness to be a property . . . *across* models. It is also worth considering the concept as it applies *within* a single model. A numerical prediction of a model is said to be robust if its value does not depend much (or at all) on variation in the value of the input parameters. . . . This type of 'internal robustness' is meaningful and can be very useful. (Orzack and Sober 1993, p. 540)

Orzack and Sober go on to warn that such internal robustness is "no sure sign of truth", but in the geomorphology examples we considered, it was never meant to be. As Michael Weisberg argues in his defense "Levins was not offering an alternative to empirical confirmation; rather, he was explaining a procedure used in conjunction with empirical confirmation in situations where one is relying on highly idealized models" (Weisberg 2006b, p. 642). Indeed this is the way geomorphologists were using robustness analyses at the second juncture – not in isolation, but as a way to help compare the predictions of the reduced complexity model to nature.

As we saw, one of the challenges facing reduced complexity models is that they are often designed for the purpose of uncovering explanatory mechanisms – not for producing quantitatively accurate predictions. Hence, it is typically not appropriate to test them by a brute comparison of their quantitative predictions with observations. The explanatory mechanisms identified in the model can be the correct fundamental mechanisms operating in nature, even if the model fails to provide quantitatively accurate predictions. Hence, robustness analyses also play an important role at this second stage of identifying those qualitative predictions or trends in the model that can appropriately be compared with observations. Thus robustness analyses, while not themselves a direct form of confirmation, can be an important step in the extended process of validating highly idealized scientific models.[4]

Acknowledgement I would like to express my deep gratitude to Brad Murray for stimulating discussions about these issues and for generously sharing his expertise with me. I am also grateful to the National Science Foundation, grant 1230676, for their support of this research.

[4] To be clear, I fully accept the point cogently made by Oreskes et al. (1994) that models can never be "verified" or proven true, and that confirmation is inherently partial. I am using model "validation" in the looser sense of establishing that the model is *acceptable for the purposes for which it is being deployed given our best available scientific evidence*.

References

Beven, K. (1996). Equifinality and uncertainty in geomorphological modeling. In B. Rhoads & C. Thorn (Eds.), *The scientific nature of geomorphology: Proceedings of the 27th Binghampton symposium in geomorphology, September 27–29, 1996* (pp. 289–313). New Jersey: Wiley.

Bokulich, A. (2008). *Reexamining the quantum-classical relation: Beyond reductionism and pluralism*. Cambridge: Cambridge University Press.

Bokulich, A. (2011). How scientific models can explain. *Synthese, 180*, 33–45.

Bokulich, A. (2012). Distinguishing explanatory from non-explanatory fictions. *Philosophy of Science (Proceedings), 79*(5), 725–737.

Coulthard, T., Hicks, D., & Van De Wiel, M. (2007). Cellular modeling of river catchments and reaches: Advantages, limitations, and prospects. *Geomorphology, 90*, 192–207.

Giere, R. (2001). The nature and function of models. *Behavioral and Brain Sciences, 24*(6), 1060.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421–431.

Matthewson, J., & Weisberg, M. (2008). The structure of tradeoffs in model building. *Synthese, 170*(1), 169–190.

Murray, A. B. (2003). Contrasting the goals, strategies, and predictions associated with simplified numerical models and detailed simulations. In P. Wilcock & R. Iverson (Eds.), *Prediction in geomorphology* (pp. 151–165). Washington, DC: American Geophysical Union.

Murray, A. B., & Paola, C. (1994). A cellular model of braided rivers. *Nature, 371*, 54–57.

Murray, A. B., & Paola, C. (2003). Modeling the effect of vegetation on channel pattern in bedload rivers. *Earth Surface Processes and Landforms, 28*, 131–143.

Murray, A. B., & Reydellet, G. (2001). A rip current model based on a hypothesized wave/current interaction. *Journal of Coastal Research, 17*(3), 517–530.

Murray, A. B., LeBars, M., & Guillon, C. (2003). Tests of a new hypothesis for non-bathymetrically driven rip currents. *Journal of Coastal Research, 19*(2), 269–277.

Nicholas, A. (2010). Reduced-Complexity Modeling of free bar morphodynamics in Alluvial channels. *Journal of Geophysical Research, 115*, F04021.

Nicholas, A., & Quine, T. (2007). Crossing the divide: Representation of channels and processes in reduced-complexity river models at reach and landscape scales. *Geomorphology, 90*, 318–339.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science, 263*, 641–646.

Orzack, S., & Sober, E. (1993). A critical assessment of Levins's The strategy of model building in population biology (1966). *The Quarterly Review of Biology, 68*(4), 533–546.

Paola, C. (2001). Modeling stream braiding over a range of scales. In M. Mosley (Ed.), *Gravel-bed rivers V* (pp. 11–38). Wellington: New Zealand Hydrological Society.

Weisberg, M. (2006a). Robustness analysis. *Philosophy of Science, 73*, 730–742.

Weisberg, M. (2006b). Forty years of 'The strategy': Levins on model building and idealization. *Biology and Philosophy, 21*(5), 623–645.

Werner, B. T. (1999). Complexity in natural landform patterns. *Science, 284*, 102–104.

Wolfram, S. (1984). Universality and complexity in cellular automata. *Physica D, 10*, 1–35.

Part III
Philosophy of Science: Realism, Anti-realism and Special Science Laws

The Ultimate Argument Against Convergent Realism and Structural Realism: The Impasse Objection

Paul Hoyningen-Huene

Abstract The target of the impasse objection is any kind of scientific realism that bases its plausibility on the stable presence of some X in a sequence of theories. For instance, if X is a set of theoretical entities that remains stable even over some scientific revolutions, this may be taken as support for convergent scientific realism about entities. Likewise, if X is a similarly stable set of structures of theories, this may be taken as support for (convergent) structural realism. The impasse objection states that the conceded stability of X could also be due to the existence of an empirically extremely successful though ontologically significantly false theory. In this case, the inference from the stability of X to the probable reality of X would become invalid. The paper closes with a discussion of several counter-objections to the impasse objection.

1 The Targets of the Impasse Objection

The argument that I shall present in this paper concerns some, but certainly not all sub-positions of scientific realism. First, it does not concern plain scientific realism that states that our best mature scientific theories are just true with respect to the postulated theoretical entities and their properties. Second, it does not concern any form of entity realism in which manipulability is the main resource for claims to reality. Third, it does not concern all forms of structural realism that either bracket the general defense of realism or do not use the "structural continuity

P. Hoyningen-Huene (✉)
Institute of Philosophy, Leibniz Universität Hannover, Im Moore 21,
D – 30167 Hannover, Germany
e-mail: hoyningen@ww.uni-hannover.de

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_11,

claim" in its defense (I shall explain further below what I mean by the "structural continuity claim"). By contrast, the argument presented in this paper firstly concerns convergent scientific realism about entities (CSRE). Secondly, it concerns all forms of structural realism (SR) that do base their plausibility on the "structural continuity claim". Finally, in more general terms, it concerns any form of realism about X that bases its plausibility on the continuous presence of X in a sequence of theories. For convergent scientific realism about entities, X would be some theoretical entities; in structural realism, X would be some structures. In addition, X could be, e.g., some properties.

Let me first deal with convergent scientific realism about entities (CSRE). It is a doctrine (or rather a family of doctrines) that roughly states that we are fairly safe to make the following two core assumptions (compare, e.g., (Sankey 2004)):

1. Accepted mature scientific theories are approximately true, which means in particular that the theoretical entities postulated by them really exist (e.g., electrons, quarks, fields, big bang, selection pressures, continental plates, etc.).
2. Scientific statements about the properties of these unobservable entities become more and more accurate in the course of scientific development.

The following assumption may be seen as optional, although it is part of the name "convergent scientific realism about entities":

3. In the course of scientific development, the sequence of accepted mature scientific theories converges to a true theory.

However, for several reasons one might be reluctant to embrace assumption 3, and I shall discuss some of these reasons further below.

In the following, I shall drop the cumbersome clause "that we are fairly safe to make" these assumptions which I used when introducing them. This clause takes care of their general fallibility which is usually conceded by realists. I shall simply say that convergent scientific realism about entities makes the two above-mentioned assumptions without meaning to deny their general fallibility.

What are the main arguments for convergent scientific realism about entities (CSRE)? All arguments start from the uncontroversial observation that since the seventeenth century, there is a successive improvement of scientific theories with respect to their empirical performance. This progress is interpreted in the sense of CSRE for the following two reasons. First, in most cases theoretical objects introduced into modern science stay there for good. Most importantly, these theoretically postulated entities usually survive subsequent changes of theory. Admittedly, there are some exceptions to this rule, but according to most adherents of CSRE, these exceptions can more or less elegantly be explained away. The claim underlying this first line of reasoning may be called the "entity continuity claim". The second main argument for CSRE is the much discussed so-called miracle argument; it is sometimes referred to as the "ultimate argument for realism". Basically, the argument states that the spectacular success of science can only be understood if science gets its theoretical postulates at least approximately right;

otherwise the success of science would be a complete miracle (and miracles usually do not exist as everyone knows). The argument becomes most persuasive if the mentioned "success of science" is further specified, namely as the repeatedly demonstrated ability of science to produce so-called use-novel predictions. "Use-novel predictions" of a theory are empirical predictions of it that were not inbuilt into it during its construction. In other words, the theory in question somehow manages to have additional empirical content that was not put into it but that it produced itself, by its own resources. What are these resources? Mainly it's theoretically postulated entities and their properties, and although their choice was motivated to account for some known empirical phenomena, they are able to account for more, sometimes even for empirical phenomena unknown at the time of the invention of the theory. The bending of starlight around the sun was such a case: not inbuilt into Einstein's General Theory of Relativity, it was nevertheless predicted by it and was somewhat later empirically confirmed to exist. Wouldn't this prediction be really miraculous if the basic assumptions of Einstein's theory were wide off the mark?

Let us now move on to a short characterization of another brand of convergent scientific realism, "structural realism" (SR; the "convergent" is usually dropped in its name). Historically, SR goes back to the early twenieth century. The more recent discussion of SR begins in 1989 with a paper by John Worrall: "Structural Realism: The Best of Both Worlds?" (Worrall 1996 [1989]). This type of SR concedes a very common counter-argument against CSRE which denies the "entity continuity claim".[1] This counter-argument has been widely used by anti-realists of various kinds. The counter-argument claims that scientific revolutions sometimes drastically change theoretically postulated entities such that the entity continuity claim of CSRE cannot be upheld. In other words, one of the most important historical pillars of CSRE collapses and with it its plausibility. Theoretical entities are thus inappropriate candidates for a realist interpretation of scientific theories. However, structural realists claim that the realist cause is nevertheless not lost as the anti-realists would have it. Instead, SR proposes that theoretical (mathematical) structures are much better candidates for a realist interpretation of scientific theories as they are much more continuous through historical change of theories. Thus, SR replaces the entity continuity claim of CSRE by the "structural continuity claim" (Votsis 2011; see also Lyre 2004, p. 664). According to structural realists, the latter claim is historically much more confirmed than the former and thus a much better basis for realism, i.e., for structural realism. As seen especially clearly in the history of physics, later theories indeed incorporate the mathematical structure of their predecessors. This is evident by, among other things, the limit relations between

[1] There is another type of SR that Holger Lyre has dubbed the "French-Ladyman-type" approach to SR, contrasting it with the "Worrall-type" approach discussed above (Lyre 2010). The French-Ladyman approach applies SR directly to concrete physical theories instead of defending it at length by general arguments at a very abstract level.

theories and their successors. Thus, according to SR there is a historically stable *structural core* in physical theories which is legitimately interpreted as reflecting reality's own structure.

For both brands of realism, for scientific realism about entities and for structural realism, it may appear attractive to assume that the sequence of theories in question indeed converges to a true theory. However, there are several problems connected with this idea. What is exactly meant by the "true theory"? How can the notion "convergence of a sequence of theories" be precisely explicated? And how can the convergence of such a sequence be claimed on the basis of a finite number of elements? The easiest way out of these difficulties is to drop the assumption that the sequence of theories indeed converges to the truth. Thus, in order to defend a particular brand of realism one may only use the "entity continuity claim", or the "structural continuity claim", or any "X continuity claim", respectively, without claiming convergence of the sequence of theories, because the continuity claim suffices for the argument. The basic idea of this argument is: what has been stable through progressive scientific development qualifies as a good candidate for being real. This is an abductive argument: a possible explanation for the presence of the stable element X in the sequence of theories is that its stability is due to its at least approximately representing an aspect of reality. Once this aspect X has been hit upon in the sequence of ever improving theories, subsequent theories won't let go of it.

2 The Impasse Objection

The objection against this kind of reasoning that I am going to state exploits the principal weakness of the abductive argument just given. The argument is independent of what sort of X has been chosen; especially, it both applies to convergent scientific realism about entities and to structural realism as discussed above (Worrall-type). Let us assume that the sequence of theories in question indeed converges to some limit theory. In spite of the tremendous difficulties to spell out exactly what that means, as discussed above, the assumed convergence is certainly not logically impossible. In this case, the realists (of the different brands in question) have to claim, possibly much against their will, that the limit theory is at least approximately true – otherwise the inference to the stable element X in the sequence as being approximately true would collapse. However, there is a possibility that the realists must at least make less plausible than its opposite: that the limit theory is a *fundamentally false* theory that is capable of making *very accurate predictions*. "Fundamentally false" may mean different things to different sorts of realists. For the defender of CSRE it would mean that the entities postulated by the limit theory would be so different from the real entities (described by the true theory) that the limit theory's entities could not count as approximations to the real entities. In other words: At least some of the terms of the limit theory do not refer to the real entities. For the defender of SR "fundamentally false" would mean that the structure of the limit theory is so different from the structure of the true theory that it would be impossible to say that the limit theory's structure is preserved in the true theory's structure. That the limit theory is capable of making very accurate predictions could mean, for instance, that all its quantitative empirical predictions are correct with a relative accuracy of 10^{-100}. In other words, the limit theory's predictions would be roughly 90 orders of magnitude more accurate than the predictions of the best mature physical theories available today.

The limit theory to which the sequence of theories converges would therefore be an impasse from which it would be impossible to get at the true theory by further gradual improvements, keeping the basic entities stable, or finding a new structure of which the old structure is a special case, respectively. The objection is thus that the limit theory could be fundamentally different from the true theory. As we have no means whatsoever to say anything substantial about the limit theory apart from the fact that it is the limit of the sequence of theories in question, we have no means to decide whether or not the limit theory is indeed an impasse – ontologically or structurally far away from the true theory. In this case, the stability of X (entities, structures, or whatever) in the sequence of theories is not a reliable indicator that these theories get X approximately right, justifying the pertinent realism. This is the impasse objection to all forms of convergent realism which base their realism on the stability of X in the sequence of theories. Note especially that the apparent advantage of structural realism over entity realism, namely, that structures are much more stable than entities in the course of scientific development (see, e.g., Worrall 1996 [1989]; Ladyman 2009), is of no help against the impasse objection.

3 Counter-Objections to the Impasse Objection

There is a variety of counter-objections to the impasse objection. Their individual plausibility may strongly depend upon one's own philosophical position. If any of the counter-objections looks silly to someone and not worthy of any serious discussion, I can only apologize; someone else may assess it differently. I shall discuss five counter-objections without judging their strength.

1. A realist may concede the logical correctness of the impasse objection. However, she may hold that the stability of X (theoretical entities, structures, or anything) in the *sequence* of theories, i.e. in *several* theories, especially over revolutionary divides, is somehow more noteworthy than reflected in the impasse objection. After all, the sequence of theories in question so far culminates in our best mature theories. It is thus more likely that the stable X in the sequence of theories *is* real and that therefore, the limit theory is indeed a (or the) true theory.

The counter-objection is that this line of reasoning is fallacious.[2] The stability of entities in the elements of the sequence of theories is only a reflection of the fact *that* the sequence converges and of nothing else. The stability is thus no indicator of the approximate truth of the limit theory.

2. Could not the no-miracles-argument overcome the difficulties posed to realism by the impasse objection? The no-miracles-argument, sometimes called "the ultimate argument" for realism (van Fraassen 1980, p. 39; Musgrave 1988), basically states that the most likely, perhaps even the only explanation for the success of science is realism about non-observables. Following this line of reasoning, the miracle argument would state that the incredible success of the limit theory, say, as in the example above, a relative accuracy of 10^{-100} in empirical predictions, would be a miracle if this theory were not very close to the true theory. As there are no miracles, it is extremely likely that the limit theory is at least approximately true.

The counter-objection to this line of reasoning is that it uses a form of the miracle argument that is too unsophisticated, equating the success of science with unqualified predictive success. As the recent discussion of the no-miracles-argument has shown that it works, if it works at all (which has been doubted by several authors for different reasons, see among many others, e.g. (Frost-Arnold 2010; Hoyningen-Huene 2011), only on the basis of *use-novel predictions* that a theory produces (Worrall 1985, 1989, pp. 148–149; Carrier 1991, pp. 26–28; Earman 1992, pp. 114–115; Leplin 1997; Psillos 1999, p. 106, 2006, p. 133). As I pointed out earlier, "use-novel predictions" were not used in the construction of the theory in question such that it comes as a surprise that the theory is capable of making them, suggesting that this is due to the theory getting something fundamentally right. However, we do not know at all whether the limit theory is capable of producing use-novel predictions; there is no indication whatsoever that the limit theory will be capable of making use-novel predictions. Therefore, the miracle argument does not help to establish that the limit theory is at least approximately true – it does not apply to the limit theory, even according to its defenders, and does therefore not eliminate the impasse objection.

3. Perhaps a different application of the miracle argument could refute the impasse objection. Let us assume that in the sequence of theories, there is a theory that admits of use-novel predictions (as indeed many physical theories do). Then the

[2] This is a situation that also occurs in the sciences. For instance, in the mid 1970s there was a variety of apparently different two-dimensional lattice models that agreed in their predictions of certain crucial thermodynamic properties. Therefore, these predictions appeared to be model independent and thus especially trustworthy. However, at a conference in 1977, the Australian physicist Rodney J. Baxter presented a model that showed that most of the current models were special cases of his own more general model (see Baxter 1977). Consequently, the confidence in the model-independency of the predictions due to their production by apparently different models immediately collapsed.

miracle argument can be applied to this theory. Granting for the moment the validity of the miracle argument, we get an abductive argument that this theory is probably approximately true. If this particular theory in the sequence of theories is approximately true, then also all its successor theories will be approximately true because they represent gradual improvements of it. If all its successor theories are approximately true, then also the limit theory of this sequence (if existing) is approximately true.

The counter-objection to this application of the miracle argument runs as follows. Given the assumption that the sequence of theories converges to a limit theory, there is an alternative explanation for the capability of a theory in the sequence to produce use-novel predictions. All predictive power of the theories in the sequence, including the perhaps surprising capability of producing use-novel predictions, is explained by their convergence to the empirically extremely successful limit theory. The only property of the limit theory that is relevant for this explanation is its extreme empirical success. Thus, this explanation for the capability of a theory in the sequence to produce use-novel predictions is independent of whether the limit theory is fundamentally false or approximately true. Therefore, from the capability of a theory in the sequence to produce use-novel predictions nothing can be inferred about the truth or falsity of the limit theory. Therefore, it is also completely open whether a theory in the sequence that admits of use-novelty predictions is approximately true.

4. Due to the radical nature of the impasse objection, one may be tempted to neutralize it by assimilating it to extremely general and fundamental skeptical arguments like Cartesian doubt or doubt about the existence of an external world. According to this line of reasoning, the impasse objection presents only a logical possibility and is not really a serious argument; it derives from a fundamentally skeptical stance. Fundamental skepticism is always a logical possibility and cannot be refuted. However, fundamental skepticism is sterile and should be dismissed. Therefore, also the impasse objection should be dismissed.

The counter-objection to this line of reasoning refutes the supposition that the impasse objection derives from a fundamentally skeptical stance. The impact objection has the form of an absolutely normal mathematical argument. If someone claims that some mathematical object O has property F, this claim can be challenged by demonstrating that O may have the property non-F. In our case, the mathematical object O is the converging sequence of theories. The claimed property F of O is that the limit theory is at least approximately true. The impasse objection doubts that and shows that the limit theory could also be fundamentally false. Thus, the impasse objection objects to the very specific transition from the (conceded) fact of convergence to a property of the limit, namely, to be an approximately true theory. The impasse objection specifically states that this is a *non sequitur*. It thus belongs to a category of very specific arguments different from the class of very general skeptical arguments.

5. In the impasse objection, the burden of proof is illegitimately shifted. It is not the (CSRE, SR, or X) realist who has to show that the limit theory is at least approximately true. On the contrary, it is the opponent of realism who has to establish that the limit theory is not at least approximately true.

The counter-objection to this line of reasoning appears to be fairly clear, although it is admitted that in general the burden of proof issue is rather thorny. In our case, it is clear that the realist claims something more specific than the opponent, namely that the limit theory is at least approximately true. The opponent only claims that the limit theory is *either* at least approximately true *or* radically false. In other words, the opponent only contends that the possibility of a radically false limit theory has not been excluded. It seems obvious that the more specific claim must be argued. Here is a very similar case. If I claim that the limit of some converging sequence of numbers is between 1 and 10, and you claim that the limit is 5, then you must justify your more specific claim. The case can also be made on the basis of the counter-objection to the third objection, above. The opponent claims that in general the inference from the existence of a limit theory to a specific property of the limit theory (approximate truth) is not valid whereas the realist claims that in the particular given case it is. It is then the realist who has to present an argument why in the particular case the inference is indeed valid.

4 Conclusion

An often used core argument supporting various kinds of scientific realism is the stability of some X (theoretical entities, structures, or whatever) in the historical sequence of theories. The impasse objection states that this continuity could also be produced by a fundamentally false but empirically very accurate limit theory. If this is correct, then the stability of X in the historical sequence of theories is not an indicator of X's representing something real, and does thus not support the respective kind of realism. Even after the discussion and, hopefully, refutation of five counter-objections to the impasse objection, I cannot claim that there are not other and possibly much stronger counter-objections. Therefore, it is certainly not excluded that the given "ultimate" argument against a specific support of some kinds of realism will share the fate of other supposedly ultimate arguments, namely, to be quite transitory.

Acknowledgements I wish to especially thank Thomas Reydon for inspiring remarks on an earlier draft of this paper, but also Claus Beisbart, Matteo Collodel, Nils Hoppe, Simon Lohse, Holger Lyre, Eric Oberheim and an anonymous referee for insightful and constructive comments.

References

Baxter, R. J. (1977). Soluble models on the triangular and other lattices. In D. Cabib, C. G. Kuper, & I. Riess (Eds.), *Annals of the Israel Physical Society. Statistical physics, statphys 13, Proceedings of the 13th IUPAP conference held 24–30 August, 1977 at the Technion Israel Institute of Technology* (Vol. 2, pp. 37–47). Bristol: Hilger.

Carrier, M. (1991). What is wrong with the miracle argument? *Studies in the History and Philosophy of Science, 22*, 23–36.

Earman, J. (1992). *Bayes or Bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

Frost-Arnold, G. (2010). The no-miracles argument for realism: Inference to an unacceptable explanation. *Philosophy of Science, 77*(1), 35–58.

Hoyningen-Huene, P. (2011). Reconsidering the miracle argument on the supposition of transient underdetermination. *Synthese, 180*(2), 173–187.

Ladyman, J. (2009). Structural realism. In Edwin N. Zalta (Ed.), *Stanford encyclopedia of philosophy.* http://plato.stanford.edu/entries/structural-realism/

Leplin, J. (1997). *A novel defense of scientific realism*. Oxford: Oxford University Press.

Lyre, H. (2004). Holism and structuralism in U(1) gauge theory. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 35*(4), 643–670.

Lyre, H. (2010). Humean perspectives on structural realism. In F. Stadler (Ed.), *The present situation in the philosophy of science* (pp. 381–397). Dordrecht: Springer.

Musgrave, A. (1988). The ultimate argument for scientific realism. In R. Nola (Ed.), *Relativism and realism in science*. Dordrecht: Kluwer.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. London: Routledge.

Psillos, S. (2006). Thinking about the ultimate argument for realism. In C. Cheyne & J. Worrall (Eds.), *Rationality and reality: Conversations with Alan Musgrave* (pp. 133–156). Berlin: Springer.

Sankey, H. (2004). Scientific realism: An elaboration and a defence. In M. Carrier, J. Roggenhofer, G. Küppers, & P. Blanchard (Eds.), *Knowledge and the world: Challenges beyond the science wars* (pp. 55–80). Berlin: Springer.

van Fraassen, B. C. (1980). *The scientific image*. Oxford: Clarendon.

Votsis, I. (2011). Structural realism: Continuity and its limits. In P. Bokulich & A. Bokulich (Eds.), *Scientific structuralism* (pp. 105–117). Dordrecht: Springer. Available at http://philsci-archive.pitt.edu/5233/1/VotsisStructuralRealismContinuityanditsLimits.pdf

Worrall, J. (1985). Scientific discovery and theory-confirmation. In J. C. Pitt (Ed.), *Change and progress in modern science* (pp. 301–332). Dordrecht: Reidel.

Worrall, J. (1989). Fresnel, Poisson, and the white spot: The role of successful predictions in the acceptance of scientific theories. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The use of experiment. Studies in the Natural Sciences* (pp. 135–157). Cambridge: Cambridge University Press.

Worrall, J. (1996 [1989]). Structural realism: The best of both worlds? In D. Papineau (Ed.), *The Philosophy of science* (pp. 139–165) (originally in *Dialectica, 143*, 199–124 (1989)). Oxford: Oxford University Press.

Doing Away with the No Miracles Argument

Simon Fitzpatrick

Abstract The recent debate surrounding scientific realism has largely focused on the “no miracles” argument (NMA). Indeed, it seems that most contemporary realists and anti-realists have tied the case for realism to the adequacy of this argument. I argue that it is mistake for realists to let the debate be framed in this way. Realists would be well advised to abandon the NMA altogether and pursue an alternative strategy, which I call the “local strategy”.

1 Introduction

The main sticking point in the contemporary debate over scientific realism is the realist’s *epistemic optimism* about current science: that we are warranted in believing that our current best tested theories are at least approximately true, and so provide a broadly correct description of the observable and (crucially) unobservable features of a mind-independent world.

The standard realist argument for such optimism is the famous “no miracles” argument (NMA), which asserts that the best explanation for remarkable empirical success of our current best theories – e.g. their novel predictive accuracy and instrumental utility as background theories (producing successful experiments, methodologies, and so forth) – is that their central theoretical claims, including those

S. Fitzpatrick (✉)
Department of Philosophy, John Carroll University, 20700 North Park Boulevard,
University Heights, OH 44118, USA
e-mail: sfitzpatrick@jcu.edu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_12,

concerning unobservable entities and structures, are at least approximately correct. Indeed, it seems that most contemporary realists and anti-realists have tied the case for realism to the adequacy of this argument. Recent discussion has therefore focused on the dialogue between the NMA and Laudan's (1981) equally famous objection from the pessimistic induction (PI), which holds that since the history of science is a wasteland of similarly successful yet now abandoned theories, we have no grounds for asserting that empirical success is best explained by approximate truth – indeed, we have more reason to expect that current successful theories will face a similar fate.

This dialogue has shaped the terms of the debate such that realism is generally taken to stand or fall with the following pair of claims about the history of science, judged necessary to retain the putative link between success and truth:

Approximate truth claim (AT): Most genuinely empirically successful yet superseded theories can in fact (from the perspective of current science) be regarded as approximately true in some substantial respect.

Continuity claim (C): The theoretical constituents that fueled the genuine empirical success of these superseded theories are, in one way or another, retained by our current best theories in the relevant domain.

This framing is evident in the discussions of prominent realists (e.g. Psillos 1999), structural realists (e.g. Worrall 1989), and anti-realists (e.g. Stanford 2006) alike.

In this paper, I will argue that it is mistake for realists to let the debate be framed in this way. Not only is the NMA far from essential to the defence of realism, it actually weakens rather than strengthens the realist cause. Instead, realists would be well advised to pursue an alternative strategy, which I call the "local strategy". This strategy has a number of significant advantages over the NMA – in particular, it has a much easier time with the PI. Realists should therefore do away with the NMA.[1]

2 An Alternative Strategy

Instead of mounting a global argument for realism like the NMA, which is intended to support epistemic optimism about all current highly successful theories in mature sciences, without our having to consider the details of these theories individually,

[1] My thesis is similar to that of Magnus and Callender (2004), who argue that the realism debate should focus on "retail" rather than "wholesale" arguments (the NMA and the PI being instances of wholesale arguments). However, their position is based on the claim (due to Colin Howson and Peter Lewis respectively) that the NMA and the PI commit the base-rate fallacy. While I agree with this move towards retail arguments, I do not think that the charge of fallaciousness will stick: Psillos (2006) shows that it depends on inappropriate probabilistic formulations of the arguments. I will show that the retail/local strategy should be preferred to the NMA, even if the NMA and PI are assumed to be non-fallacious.

this alternative strategy says that the defence of realism is best constructed on a case-by-case basis. The idea is that the best foundation for a realist attitude towards a particular theoretical claim of modern science (e.g. that there are atoms, that past and present organisms on earth are the product of evolution by natural selection, that the continents move laterally on tectonic plates, etc.) is the weight of the particular first-order evidence that led scientists to accept the claim in the first place. Realism is thus to be defended through close consideration of the specific theoretical claims that realists want to be realists about, the particular empirical evidence for such claims, and questions about what epistemic attitude towards these claims is licensed by this evidence, with anti-realist challenges to be rebutted as they arise.

A good example of this sort of strategy is Achinstein's (2002) discussion of Jean Perrin's early twentieth century arguments for the atomic theory, which posits that chemical substances are composed of unobservable atoms and molecules, and that Avogadro's number, N (the number of molecules in a sample of a substance whose weight in grams is equal to the molecular weight of the substance), is a constant approximately equal to 6×10^{23}. Achinstein (following an earlier suggestion from Wesley Salmon) claims that Perrin's experiments on Brownian motion provide a compelling case for realism about the atomic theory – indeed, one that convinced many of Perrin's anti-realist contemporaries. Combined with those of Gouy and others, they eliminated all admissible causes of Brownian motion other than internal molecular forces, and, by providing multiple convergent estimates of N, Perrin provided quantitative confirmation of the atomic structure of matter. Close consideration of Perrin's arguments, Achinstein argues, demonstrates that standard anti-realist objections to epistemic optimism about the theory are misplaced or rendered ad hoc. In particular, he is concerned to show that a van Fraassen-style agnostic empiricism about the existence of molecules is hard to maintain, given Perrin's evidence, without resorting to arguments for full-blown inductive or general epistemic scepticism – yet van Fraassen (like most contemporary anti-realists) claims only to be a *selective* sceptic about claims about unobservables.

Based on this case study, Achinstein claims that there can be a valid experimental argument for realism about a theory. The suggestion seems to be that similar local experimental arguments can be articulated for other key components of the contemporary scientific picture; hence, realists should pursue a similarly piecemeal strategy elsewhere.

Though Achinstein points realists in the right direction, he is wrong to suggest that the argument for realism in this case is purely experimental.[2] Realism is, after all, a *philosophical* position about science, and modern anti-realists do not see themselves as rejecting the evidence that scientists advance for their theories, but rather as arguing for the correct philosophical interpretation of its epistemic significance. The philosophical action in Achinstein's discussion takes place in his rebuttal of various anti-realist interpretations of Perrin's theory and evidence (such as van Fraassen's), which Perrin himself did not consider, rather than in

[2] Saatsi (2010) and Psillos (2011b) make a similar point.

the recitation of Perrin's evidence itself. Moreover, not all of the considerations Achinstein invokes are specific to this case. For example, he argues that we have empirical reasons to think it ad hoc to be sceptical about inferences from properties of observed bodies (e.g. visible Brownian particles) to properties of unobservable bodies (e.g. molecules), but not about inferences from properties of observed bodies to properties of observable but unobserved bodies, given that we can vary the features in virtue of which bodies are observable (e.g. size) and find that this makes no difference to the properties of interest (e.g. that bodies have mass), and we have no other grounds for taking unobservability to be a biasing condition.

Thus, as I characterise it, the local strategy is not simply a matter of reciting first-order evidence for the particular theoretical claim at hand, since it will require philosophical argumentation that goes over and above this evidence, and consideration of issues that may transcend the local details of specific cases.[3] However, the strategy does hold that the details of specific cases are to be centre stage. Achinstein's key insight, I think, is that it is much easier to run standard anti-realist arguments and adopt anti-realist views of current science at a high level of abstraction, when one doesn't have to consider specific cases like Perrin's in any detail. Though first-order evidence will not be sufficient to establish realism by itself, such details nonetheless provide the realist with powerful resources to support her position. Thus, from a realist perspective, the battle is best fought in close contact with such evidence, rather than on the global level where the NMA operates, where such details are glossed over.

Clearly, much more needs to be said to give a full characterisation of the local strategy, and to show that it is in fact capable of defending a realist attitude towards particular theoretical claims. My concern here, however, is to highlight the strengths of this approach relative to the NMA. I will begin by highlighting its advantages when it comes to dealing with the PI.

3 Diffusing the Pessimistic Induction

The NMA provides easy grist for the mill of the historically motivated pessimist because it asserts a completely general connection between empirical success and approximate truth. A list of successful yet false theories from the history of science would call this into question. The NMA-focused realist must therefore offer a defence of sweeping historical claims AT and C in order to maintain the putative link between success and truth. Whatever one's suspicions about the tenability of these claims, this is clearly a burdensome task for the realist to take on.

[3] As an example of the latter, there is the question of what kind of evidence will be sufficient to warrant a realist attitude? This is clearly a general philosophical question, the answer to which hangs on one's *theory* of evidence, rather than some particular evidence itself.

It is much harder, however, to formulate a PI that threatens any particular instance of the local strategy. The most obvious way to try to do so is to point out that past scientists thought they had very strong evidence for theoretical claims – the existence of the ether, phlogiston, and so forth – now judged to be false; hence, we have grounds for pessimism about current theoretical claims held to be strongly supported by evidence. But, as Roush (2010) has argued, in an insightful piece on the PI, that kind of induction obscures the fact that the *content* of the evidence we have for current theoretical claims is quite different to that which supported now abandoned claims of previous scientists. The particular experimental results appealed to by Perrin, for instance, are quite different to those appealed to by ether or phlogiston theorists, and warrant quite different theoretical conclusions. Importantly, the realist needn't deny that past scientists had (good) evidence for their claims. Since one cannot draw a conclusion about apples from an induction over oranges, the realist can undermine a PI over first-order evidence just by pointing out that the relevant scientists whose work forms the basis of the particular instance of the local strategy built their claims on *different* sets of evidence.[4] The burden is then shifted onto the pessimist to explain why these evidential differences make no *epistemic* difference.

Clearly, the way pessimist will respond is to move to the meta-level: claims about ether and phlogiston may have been based on different sets of evidence to those that underpin Perrin's claims, but the scientists concerned presumably used the same inferential methods in arriving at such claims. Hence, a history of inferential failures motivates pessimism about the reliability of the methods that scientists (including Perrin) use in apportioning confidence in theoretical claims given the available evidence, whatever its content.

In response to such a meta-level PI, Roush argues that there has in fact been very significant methodological change over the course of recent scientific history. Experimental and statistical methods, for instance, have changed drastically in the last century, and much of this change has been driven by the recognition of past methodological failures. Hence, Roush claims that current scientists plausibly go about apportioning confidence in particular theories based on the available evidence in ways that are significantly different to those of previous scientists, again shifting the burden onto the pessimist to explain why the failures of past methods deployed by past scientists are epistemically relevant to confidence in theoretical claims that are highly warranted according to current methods, given the evidence available to us now.

One concern about this response is that if there have been such drastic changes in method, we might anticipate equally significant methodological changes in the future, undermining confidence in the reliability of current methods.[5] Moreover,

[4] Within the domains once occupied by ether and phlogiston theories, successor theories are presumably also grounded in different evidence sets, including, for instance, the evidence that led to the *rejection* of their predecessors.

[5] I thank a reviewer for raising this objection.

much of the realism debate has been concerned with the reliability of inference to the best explanation (IBE), and one could argue that, at least at some general level, this sort of inferential method has had a relatively stable presence in the armoury of scientists (whether they have been aware of it or not), particularly when justifying claims about unobservables.

A better response, in my view, which is more in keeping with the spirit of the local strategy, is to emphasise the *contextual* nature of scientific inference. Even if, at some general level, current scientists use the same, or very similar inferential methods to those of past scientists, how those methods get deployed in particular cases depends heavily on the local scientific context. This context is crucial for understanding the justification for the particular inferences scientists make. Though Perrin's argument from his multiple convergent estimations of N could perhaps be viewed as an IBE (the best explanation for this concordance being the existence of molecules) – hence, methodologically like, say, inferences to the existence of the ether – as Achinstein's discussion makes clear, the epistemic force of this argument can only be appreciated within the context of a rich network of background beliefs: for example, Perrin's belief that his and others' experiments had ruled out conceivable causes of Brownian motion other than internal forces, that visible Brownian particles behave like invisible molecules and so could be used to calculate N, that experiments on phenomena other than Brownian motion had reached similar values for N, and that such values for N are not to be expected on the denial of the atomic theory. It is these local background beliefs that underwrite Perrin's inference to the reality of molecules, given his results. Anti-realists may, of course, take issue with these beliefs, or question whether they do in fact license this inference, but these are not questions to which the historical track record of IBE in quite different contexts seems directly relevant.

The way the proponent of the local strategy can respond to the meta-level PI, therefore, is to argue that the grounds for epistemic optimism about a particular theory is crucially contingent on context-specific background beliefs. It is misguided, therefore, to lump together scientists' inferences from evidence in a context-independent fashion, talking about the reliability or unreliability of inferential methods such as IBE quite generally, based on their historical track record, since that obscures the particular epistemic features that underwrite them.[6]

This brief discussion is certainly not a fully adequate response to the potential historical objections to the local strategy.[7] My aim, however, has merely been to

[6] As Saatsi (2010) points out, Achinstein's content-driven approach has some affinity with Norton's (2003) material theory of induction. According to Norton, what makes particular inductive inferences justified or unjustified is not their conformity to some formal inference schema (e.g. their being IBEs), but rather local material facts believed to obtain in the domain of inquiry. It should be noted that Achinstein himself (2010, Chap. 4) does not accept Norton's claim that there are no valid universal rules of induction, but does accept that local background information is crucial in understanding the license for particular applications of inductive rules.

[7] Stanford's (2006) new formulation of the PI also has to be considered. This trades on the apparent inability of past scientists to conceive of all the conceivable alternative hypotheses equally or better supported by the available evidence, since they frequently failed to conceive

make the case that this strategy has more powerful resources for dealing with such objections than the NMA. The local strategy invites us to pay close attention to the content of particular scientific theories and inferences, allowing realists (hopefully) to highlight relevant local features that can ground epistemic optimism, and which distinguish the theoretical claims and inferences at hand from those of past scientists. In contrast, the generality of the NMA makes it too easy for the anti-realist to invoke the history of science in the case against realism, thus encumbering the realist with the burden of defending sweeping historical claims like AT and C. Advocates of the local strategy do have to delve into the details of current science and highlight relevant epistemic differences between particular cases of current science and past science in order to fend off historically motivated pessimism. However, nothing so burdensome as AT and C need be at stake here.

4 Defending the Realist Framework

In recent work, Psillos, one of the most ardent defenders of the NMA, has expressed much sympathy with Achinstein's claims about Perrin – indeed, claiming that "Perrin's case is so strong that the first-order evidence for the reality of molecules takes precedent over the second-order evidence [e.g. from the PI] there might be for being sceptical about explanatory posits" (2011b, p. 188). However, he thinks that such local arguments for realism, while important, are necessarily incomplete. He cites two reasons for this. I'll discuss the first here, the second in the following section.

Psillos (2011b) argues that close consideration of cases such as Perrin's fail to establish what he calls the "realist framework". This is a general philosophical framework, which *allows* that we can explain observable phenomena by positing the existence of theoretical entities and structures (e.g. as micro-constituents of observable phenomena), and that claims about such entities and structures can, in principle, be confirmed by empirical observations – e.g. in virtue of their explanatory role. Perrin clearly assumed such a framework, since he thought that it was possible to construct legitimate arguments from empirical observation to the reality of molecules. But, it is hard to see how his arguments can have any force against someone who simply rejects such a framework. It is notable that, though Perrin was able to convince many anti-realist opponents of the atomic theory, there were holdouts like Duhem, who never accepted the reality of atoms and molecules. Duhem (1991) regarded the positing of such unobservable entities as dubious and unnecessary "metaphysics"; hence, he was never prepared to accept the legitimacy of inferences like Perrin's from observation to the existence of molecules. It seems, then, that no amount of close consideration of the available

of alternatives that subsequently came to be regarded as better confirmed by that evidence. Since such cognitive limitations presumably also affect current scientists, this motivates pessimism about current theories.

"evidence" for theoretical entities such as molecules is likely to settle the general question of the legitimacy of explanation by postulation of theoretical entities and structures, and whether or not such things can, in principle, be confirmed. Any instance of the local strategy must *presume*, but cannot independently establish, the plausibility of this realist frame.

Historically, part of the seeming attraction of the NMA has been the idea that it might have some independent pull against such hard-line anti-realisms. This is because it works with a purely instrumental notion of empirical success – novel predictive accuracy and instrumental utility *qua* background theory – that even a Duhemian instrumentalist can accept (e.g. Boyd 1984, p. 59). However, as Psillos (2011a, c) is now happy to concede, the very notion that we need to explain the instrumental success of theories, and that such successes can support an attitude of epistemic optimism towards claims about theoretical entities and structures is precisely what is at issue in this question of framework. A Duhemian anti-realist can very well accept the instrumental success of theories, but stop short at the inference from success to (approximate) truth, by denying the legitimacy of explanation by postulation. At worst, then, the local strategy is in the same boat as the NMA: neither strategy can provide an independent argument for the realist framework, but only for certain realist positions *within* that framework.

What this shows is that realists need different arguments for different realist claims. One argument (or set of arguments) will be required for the general realist framework – for example, Psillos (2011a) has suggested an interesting argument for the indispensability of theoretical entities based on the ideas of Schlick, Feigl, and Reichenbach. One possible position in that framework is a view – consistent with various forms of epistemological anti-realism – which holds that it is perfectly legitimate to posit theoretical entities and structures, but which holds that we lack sufficient warrant to believe that any current theoretical claim is approximately true. Other arguments will then be required for epistemic optimism, and that is where the local strategy comes in. Importantly, within such a framework, there is no need to confine oneself to a purely instrumental conception of the grounds for epistemic optimism.

5 So What Use Is the NMA?

The second reason that Psillos cites for the incompleteness of local realist strategies, such as Achinstein's, is that he thinks that realists also need to offer a general justification for the reliability of IBE. For Psillos, the primary payoff of the NMA *within* the realist framework is that it provides a (non-vicious) rule-circular justification for the reliability of IBE. His formulation of the argument has two parts:

(A)

(A1) Scientific methodology is theory-laden.

(A2) These theory-laden methods lead to correct predictions and experimental success (instrumental reliability). How are we to explain this?

(C1) The best explanation (of the instrumental reliability of scientific methodology) is this: the statements of the theory which assert the specific causal connections or mechanisms in virtue of which methods yield successful predictions are approximately true.

(B)

(B1/C1) Theories are approximately true.

(B2) These background scientific theories have themselves been typically arrived at by abductive reasoning.

(C2) Therefore, (it is reasonable to believe that) abductive reasoning is reliable: it tends to generate approximately true theories. (Psillos 2011c, pp. 23–24)

In contrast, I think there is good reason to doubt that the NMA can add anything of value here. Much in line with the ideas floated at the end of Sect. 3, Psillos has argued that the structure and strength of IBE reasoning is determined by local context:

> IBE-type of reasoning has a fine structure that is shaped, by and large, by the context . . . The background knowledge (or, beliefs) ranks the competitors. Other background assumptions determine the part of the logical space that we look for competitors. The relevant virtues or epistemic values are fixed, etc. Given this rich context, one can conclude, for instance, that the double-helix model is the best explanation of the relevant evidence, or that the recession of the distant stars is the best explanation of the red-shift . . . These contextual factors can link [explanatory] loveliness and likeliness nicely, because they do not try to forge an abstract connection between them; rather the connection stands or falls together with the richness and specificity of the relevant information available. (Psillos 2007, p. 443).

These sentiments do not mesh well with the idea that IBE is to be justified by a global track record argument like the above formulation of the NMA. If IBE is so context-sensitive, it does not seem appropriate to talk of the reliability of IBE in general, but only of particular instances of IBE in particular contexts. Hence, rather than ask for a general justification for IBE, it seems more appropriate to ask how it is that particular contextual information licenses particular instances of IBE. It is hard to see, therefore, what justificatory role the NMA can play.

6 Doing Away with the NMA

The NMA has been, and remains, the primary realist argument for epistemic optimism about current science. But, as we've seen, there is an alternative strategy – the local strategy – that has very significant advantages over the NMA. Most notably, it seems to have a much easier time with the PI, and doesn't encumber the realist with such onerous historical commitments. The local strategy leaves untouched the question of what Psillos has called the "realist framework", but then so does the NMA. In addition, the NMA does not seem well suited to the job that it has been thought to play within that framework: that of providing a general justification for IBE.

If realists were indeed to do away with the NMA, and instead utilise the local strategy for defending epistemic optimism about current science (leaving the question of the realist framework to other arguments), this would, I suggest, lead to a much better framing of the realist debate, for realist purposes. First, it is curious feature of the recent debate that in all the papers and books devoted to the topic one hardly sees any discussion of the particular current scientific theories that realists actually want to be realists about. In recent years, the fuss seems to have been almost exclusively over various features of the historical record – mostly a select few of the cases cited by Laudan. This is clearly an effect of the focus on the NMA. But, while I do not deny that history is relevant to the realism debate, it is surely current theories that ought to be the focus of attention. The local strategy thus focuses the debate where it should be: on the details of specific current theories, the specific evidence for them, and questions about what epistemic attitude towards the entities and structures postulated by these theories is warranted by this evidence.

Second, the local strategy helps to put the scope and limits of realist optimism into proper focus: realism may be appropriate for some claims of current science, but not others, and we have to delve into the details of particular cases to find out which. In so doing, it helps move the debate away from the overgeneralised and unnecessarily concessive positions that realists have been inclined to adopt as a result of their attempts to rescue the NMA. Worrall's (1989) structural realism is a case in point. While it may perhaps be true that all we are entitled to be optimistic about in current science are its implications about the abstract structure of nature, it seems bizarre to make this as a general claim, as Worrall does, on the basis of a few cases from nineteenth century physics, where it seems that the NMA can be defended against the PI only at the level of structure. This looks at once like hasty overgeneralisation (why should the fortunes of a few theories from the history physics be taken as a model for all of current science?) and, from a realist perspective, far too concessive to the anti-realist, made as it is without actually engaging with the non-structural claims of specific current theories.[8]

Realists should therefore do away with the NMA.

Acknowledgement Thanks to the audience at EPSA11, Stathis Psillos, and a reviewer for helpful criticisms and suggestions.

References

Achinstein, P. (2002). Is there a valid experimental argument for scientific realism? *Journal of Philosophy, 99*, 470–495.

Achinstein, P. (2010). *Evidence, explanation, and realism*. New York: Oxford University Press.

Boyd, R. (1984). On the current status of scientific realism. In J. Leplin (Ed.), *Scientific realism* (pp. 41–82). Berkeley: University of California Press.

Duhem, P. (1991). *The aim and structure of physical theory*. Princeton: Princeton University Press.

[8] Structural realists have appealed to other kinds of arguments in support of the claim that all we can know about nature is its structure (see Ladyman 2009). However, the argument from theory change was primary in Worrall's original papers and it remains a focus of attention in the current literature.

Ladyman, J. (2009). Structural realism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2009 Edition). http://plato.stanford.edu/archives/sum2009/entries/structural-realism/. Accessed June 2012.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science, 48*, 19–49.

Magnus, P. D., & Callender, C. (2004). Realist ennui and the base rate fallacy. *Philosophy of Science, 71*, 320–338.

Norton, J. (2003). A material theory of induction. *Philosophy of Science, 70*, 647–670.

Psillos, S. (1999). *Scientific realism*. London: Routledge.

Psillos, S. (2006). Thinking about the ultimate argument for realism. In C. Cheyne & J. Worrall (Eds.), *Rationality and reality: Conversations with Alan Musgrave* (pp. 133–156). Dordrecht: Springer.

Psillos, S. (2007). The fine structure of inference to the best explanation. *Philosophy and Phenomenological Research, 74*, 441–448.

Psillos, S. (2011a). Choosing the realist framework. *Synthese, 180*, 301–316.

Psillos, S. (2011b). Making contact with molecules: On Perrin and Achinstein. In G. Morgan (Ed.), *Philosophy of science matters: The philosophy of Peter Achinstein* (pp. 177–190). New York: Oxford University Press.

Psillos, S. (2011c). The scope and limits of the no-miracles argument. In D. Dieks, W. J. Gonzalez, S. Hartmann, T. Uebel, & M. Weber (Eds.), *The philosophy of science in a European perspective* (Vol. II, pp. 23–35). Dordrecht: Springer.

Roush, S. (2010). Optimism about the pessimistic induction. In P. D. Magnus & J. Busch (Eds.), *New waves in philosophy of science* (pp. 29–58). Basingstoke: Palgrave Macmillan.

Saatsi, J. (2010). Form-driven vs. content-driven arguments for realism. In P. D. Magnus & J. Busch (Eds.), *New waves in philosophy of science* (pp. 8–28). Basingstoke: Palgrave Macmillan.

Stanford, K. (2006). *Exceeding our grasp*. New York: Oxford University Press.

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica, 43*, 99–124.

Theory-Parts for Scientific Realists

Alberto Cordero

Abstract The "divide and conquer" approach to scientific realism requires a compelling criterion for specifying theory-parts worthy of realist commitment – components deemed very probably true. Articulating such a criterion has proved difficult, however. Long-term retention across theory-change provides a poor basis at best (judging by examples like the ether of light); the same can be said about such features as empirical success and current freedom from specific doubts. This paper argues for a seemingly better choice, drawn from scientific practice and focused on five overlapping strategies: hostile probing of a theory's central tenets by its opponents, revision of auxiliary assumptions (mainly by supporters of the theory), external grounding of theoretical assumptions, efforts to identify adequacy conditions for successor theories, and explanations of a theory's success after becoming superseded.

1 Scientific Realism

I will use the term "scientific realism" to refer to a position that is (a) "scientific" in its focus on accounts judged successful by the toughest standards of scientific practice, and (b) "realist" in its allowance for ampliative inference into domains unreachable by unaided human perception; (c) naturalistic in its approach to warrant and justification; and (d) scientific rather than "philosophical" about truth. Also (e)

A. Cordero (✉)
Graduate Center CUNY & Queens College CUNY, City University of New York,
New York, NY, USA
e-mail: corde@prodigy.net

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_13,

the realism at play is of the “Divide and Conquer” variety (DAC) – i.e. a position that claims truth for selected theory-parts and narratives (as opposed to whole theories). This paper argues for a criterion by which theory-parts suitable for DAC-realism can be identified, including much of what current science claims about sub-atomic particles, atoms, molecules, chemical structures and material transformations, microorganisms, biological structures, , geological and evolutionary histories, along with significant portions of the accounts and dynamical narratives yielded by present theories about various aspects of the world. The proposal that follows draws from DAC realist moves made in recent decades by some thinkers, especially Philip Kitcher, Jarrett Leplin, Stathis Psillos (KLP), and Paul Thagard.[1]

According to DAC realism, it would be wrong to treat a discarded theory as merely a conjunction of hypotheses that fail just because some of its components are false. On this view, proposals now superseded, but with track records rich in corroborated novel predictions, still give correct theoretical descriptions at many levels (e.g. Newton’s gravitation theory and the ether-based theory of light), unlike theories lacking in such predictions (e.g. the crystalline spheres and the four humors).

One key task of DAC realism is thus to identify in successful theories restricted and/or coarse-grained models whose applications produce theoretical descriptions that are very likely true. We can think of these bodies as the correct “parts” of the theory. The question is what identifies a correct theory-part. As Laudan (1984) saw, retention alone is not a good marker of truth. Many long kept ideas have turned out badly: those of caloric, phlogiston, teleological holism in biology, the luminiferous ether, the Euclidean conception of space and time from Antiquity to Lorentz. Realists need a better marker.

Here is an influential proposal about epistemically promising theory-parts: although past successful theories license many false claims, assertions that are really off the mark have never been implicated in the predictions that crowned the theories in question. On this view, developed by Kitcher in the 1980s and subsequently by Leplin and Psillos, synchronic analysis of the allocation of evidential weight suffices to separate the wheat from the chaff. For example, although Fresnel’s successful theory of light explicitly appealed to the ether, he did not need to do so in order to obtain his famous experimental predictions. Regarding the latter, KLP urge, the ether concept was “idle,” “dispensable” or worse. In their view we can tell this simply by checking the theoretical options that were available in Fresnel’s time and allocate epistemic weight to theory-parts accordingly.

This “synchronic” version of DAC has received critical attention over the last decade. Commentators widely agree that these attempts at synchronic determination of sound theory-parts fail – e.g. Chang (2003), Carrier (2004), and Chakravartty (2003, 2007), Cordero (2011a, b).

How about diachronic approaches? In this paper I discuss four diachronic realist strategies and spell out their promise and limitations as realist markers applied to

[1] Kitcher (1993), Leplin (1997), Psillos (1999), and Thagard (2000, 2007).

the DAC variety. One particularly well-received strategy, based on work by Thagard and others, highlight theory-parts that gain elucidation from some ongoing external theory. I look at this strategy from a DAC perspective; one major weakness it has resides in the tricky epistemic value of explanation. I then consider two related diachronic DAC moves that build on the way William Whewell looked at epistemic success in the nineteenth century. Unfortunately this realist line falls prey to the problems similar to those faced by the KLP approach. Next I consider and assess two further strategies common in scientific theorizing; one seizes on the articulation of adequacy conditions for theory revision; the other seizes on explanation of the success of superseded theories from the vantage point of current theory. I argue that, contrary to appearances, none of these DAC strategies amount to post-hoc maneuvering; but they show other weaknesses. Finding that each of the reviewed strategies carries some positive confirmational import, yet not enough to do by itself the required job, I consider combinations of the strategies that might work. The paper closes with a suggestion of a way to pull together the promising aspects of the reviewed strategies in favor of a better criterion for theory-parts of realist interest, hard-DAC. The latter seems to fulfill the goal of realists from Whewell on, now with a focus on theory-parts.

2 Deepening and Its Shortcomings

The first realist alternative I want to discuss focuses on the support that theory-parts gain when independently successful theories explain them. Major beneficiaries of this "lateral" support include claims first introduced as either postulates or experimental findings – e.g. the explanation provided by Maxwell's theory of such features of Fresnel's theory as the speed of light and the transversal character of light waves. How much added credibility this kind of elucidation brings to a theory-part depends on how well-established the elucidating theory is. The physical and biological sciences are rich in compelling success stories in this regard. Think of the numerous assumptions of cell biology explained by molecular biochemistry – e.g. neural mechanisms accounted for by noting that neurons consist of proteins and other molecular components organized into functional areas (e.g. nucleus, mitochondria, axons, dendrites, and synapses).

One influential line of this DAC approach follows Paul Thagard's emphasis of "deepening." According to Thagard, having an account of why a mechanism postulated by a theory T works improves the explanatory value of T, adding to its credibility. On this view, if a theory not only maximizes explanatory coherence, but also broadens its evidence base over time and its assumptions receive elucidation in the form of explanations of why the theory's proposed mechanisms work, then we can reasonably conclude that the theory is at least approximately true. That is, explanatory unification with previously separate theories contributes inductive weight to the theory-parts thereby elucidated.

Theoretical elucidation or deepening has undeniably accompanied much of the advance of modern science. But how strong a marker of probable truth is the elucidation at play here? Deepening is something a theory often (if not exclusively) gets after it is superseded, as part of explanations of why the theory worked as well as it did. In the case of the classical atomic theory of matter, for example, the deepening bought on by successor theories involved abandoning the claim that atoms do not have parts: strictly speaking, the old atomic theory was not properly elucidated so much as rendered false. Still, although false as originally stated, atomic theory was "approximately correct:" at least within certain specifiable contexts, enough theoretical texture was retained for claiming that certain aggregates of sub-atomic matter behave almost exactly as the discarded theory said atoms do (Bohm 1957).

One initial complaint about this strategy goes as follows: To the extent that the history of science displays effective retention of deepened parts, it also pushes the elucidation thesis towards a weaker probabilistic connection between elucidation and approximate rather than exact truth. Another complaint is that, like many contemporary realists, Thagard takes a theory to be approximately true if "most of its claims are nearly true in achieving quantitative closeness to accepted values." In Thagard's view, in a deepened theory, something or other is likely to be approximately true, such that most of the claims about the constituents of things licensed by the theory are approximately true. The realist thesis is thus that, in a theory exalted by deepening, some of its parts will display strong retention in successor theories *because* the assertions they license are approximately true.

Unfortunately, there are problems with this way of advancing realism. As presented, elucidation and deepening are as promising for advancing realism as they are risky. One source of trouble is the noted expectation of truth for "most" of the claims issued by a successful theory. A second line of problems has to do with the excessive epistemological cheerfulness of explanationist ploys advanced on behalf of scientific realism.

The common realist claim that "most" assertions licensed by a successful theory are approximately true clashes with reality; fortunately, such a claim is unnecessary for realism. Regarding virtually all past successful theories, "most" of the claims licensed by superseded theories have turned out to be false. While there may generally be reason to believe that a successful theory T gets right some of what it says about unobservables, there is no reason to expect that successor theories will regard "most" claims T yields about anything as true (particularly at the most fundamental levels). Take, for instance, the assertions derived from classical Newtonian gravitation about, say the stones of Athens. From a realist perspective, most such assertions are utterly false, because, in terms of speeds and gravitational environments, the range of possibilities now believed to apply to the said stones reaches well outside the "Newtonian range" – the domain that Newtonian theory describes accurately enough for sustaining a realist stance.

Happily for realists, however, this difficulty seems one created by a greedy formulation of the realist hope in sight, a formulation not needed for making the relevant point. As already stressed, realism about anything like the whole or "most" of what a theory says about its intended domain has long been exposed as a dubious project in light of numerous historical examples, no matter how successful the theory at hand may be. Realists have responded by shifting their focus to theory-parts, specifically to assertions of partial models of successful theories – partial in the sense of restricting their applicability to selected parts of the originally intended aspects, domains and/or giving them "appropriately" coarse-grained representations.

Which partial models generate truthful descriptions? A vague assertion to the effect that, at "*some* level *some* of the central claims posed by a successful theory are *approximately true*," can be made without appeal to deepening or explanation. In physics and chemistry, all the successful scientific theories of the last two centuries license thick arrays of partial models (of the suggested sort) that have never ceased to be truthfully assertible. There is thus the following lame realist thesis: In a theory with corroborated novel predictive power at least some parts of its novel theoretical narrative are approximately true. To many commentators, however, this realist thesis is much too vague for serious comfort, exposing the need for a more detailed specification of theory-parts.

Which specific parts of a successful theory are worthy of realist commitment, and why? This brings us to the second highlighted issue regarding explanationist expectations: their debatably excessive epistemological cheerfulness. The expectation is that, because explanatory deepening springs from independently supported theories, it raises the epistemic probability of the assumptions and narratives it elucidates – hence its significance for realists. But, in objectivist terms, how much of an epistemic raise do elucidated theory-parts get? Numerous unsavory counterexamples seem to cause trouble.

Explanationist hopes have a particularly difficult time in disciplines that encompass little precision. The fertility of the imagination thrives in vagueness. Consider, as an illustration, the elucidations entertained by Freud and some of his followers for various assumptions of the theory of psychoanalysis.[2] Freud's theory had many plainly unjustified assumptions. An instance in point is the so-called "death instinct," an urge to die allegedly present in all living creatures. Starting with Freud, some psychoanalytic writers have sought to explain this and other basic tenets of the theory in terms of reliable, independently successful physics. Freud himself thought of instinct as a form of energy transformation. In the case of the death instinct, several followers of Freud tried to describe this instinct as an application of the 2nd law of thermodynamics, according to which in a closed system all forms of energy tend to dissipate. For a while, Freud agreed that this provided respectable grounding for his ideas. As his theory became increasingly "autonomous," however, Freud lost interest in rooting it in physics. Still, for a long time many distinguished psychoanalysts continued to wax lyrical about the epistemic import of these suggested elucidations. Thinkers outside their circle did

[2] See, for example, Sulloway (1992) on relevant works by Sigmund Freud, Franz Alexander, Sigfried Bernfeld & Sergei Feitelberg, Heinz Lichtenstein, and Leon Saul.

not take kindly to the proposed embedding in reliable physics. Appealing to physical theories without including their mathematical resources and precision in the package cheapens the elucidation strategy. Theories that allow for ample vagueness limit or even halt the epistemic benefits of deepening.

What about mathematized theories? How much of an epistemic improvement do their theory-parts get from explanatory derivation? In some cases remarkably little, it seems. Consider the following explanation, intimated by Kepler for his 2nd Law.[3] It uses as premises [A] a principle of economy to the effect that actions (influences) operate without waste in nature, [B] the Aristotelian conception of force and [C] the idea that the motive force responsible for the planetary motions proceeds from a constant emanation contributed by the rotating Sun. [A] and [C] lead to the claim that the motive force at play is confined to the planetary plane and corresponds to the same amount of total pushing being available at each distance from the Sun, distributed along the corresponding circumference, which makes accessible a local pushing of magnitude f(r) at each point of the circumference. That is, for a distance r from the Sun:

$$(2\pi \mathrm{r})\, \mathrm{f}\,(\mathrm{r}) = \text{constant. So, } \mathrm{f}\,(\mathrm{r})\ \alpha\ 1/\mathrm{r}$$

Since, according to [B], a planet's speed v is proportional to f(r) above, feeding [B] into the above equation yields:

$$\begin{gathered} \mathrm{v}\,(\mathrm{r})\, \alpha 1/\mathrm{r}, \text{ and so:} \\ (\mathrm{r})\ \ \mathrm{v}\,(\mathrm{r}) = \text{constant} \end{gathered}$$

Applying the above result to the space s travelled along the circumference, and equating the speed v with (Δs/Δt), we get:

$$\begin{aligned} &(\mathrm{r}\ \Delta \mathrm{s})\ /\Delta \mathrm{t} = \text{constant; that is:} \\ &\Delta \text{ area swept by the planet}/\Delta \mathrm{t} = \text{constant} \end{aligned}$$

We thus get the 2nd Law (a planet sweeps equal areas in equal times), now *deepened* by the derivation just given, courtesy of Aristotelian physics and a principle of frugal economy for solar action. The problem is, of course, that we think [A] and [B] are downright false theoretical premises. Once again, the elucidation exercise ends in failure, in this case despite the higher mathematization involved.

One obvious insinuation should be that explanatory deepening does not suffice for realism. But then, what (if anything) might? Given a successful theory, can we tell which parts get their object right, and if so can we do so without resorting to mere retrospective projection of current theory?

[3] Kepler (1620/1995), Davis (2003).

Explanations respond to our need to understand the world, and as such they are an integral part of the scientific quest for knowledge. Too often, however, the fertility of the human imagination leads us into serious disappointments. Precautions need to be taken. As antirealists urge, "cogent explanations" are relatively easy to obtain, unless the label "cogent" carries seriously stringent conditions. The way theorizing normally receives guidance from unacknowledged background information further compounds the situation. If realists want to improve the elucidation condition by raising its requirements, one way to go is by toughening conditions of admissibility. Many current realists thus emphasize the epistemic virtue of novel prediction, especially the thesis that, if a theory yields previously unanticipated predictions and these come out true, then we have reason to believe that some nontrivial part of the theory's associated narrative is true.[4]

Arguably, the most powerful warrant of this sort comes from psychological novelty (Worrall 1989), because such novelty protects more than other kinds against unintentionally smuggling "information" into theories. To the extent that psychologically novel predictions function as promising indicators that the theory involved somehow adds to theoretical knowledge, novel predictions that prove successful advance the realist case for a theory. We find this take on prediction already in place in the 1830s in physics and chemistry, and two decades later – through Whewell (1968) and Charles Darwin[5] (1859), as at least an ideal in biology too.

Methodologically upgraded explanations that yield corroborated novel predictions raise the epistemic likelihood of the theory-parts they illuminate. This makes encouraging news for realism, even if the import of elucidation may not extend beyond a modest realist thesis. One obvious move for elucidation realism is thus to concentrate on just those elucidations that enhance the novel predictive power of the theory-parts they deepen (strong elucidation). The realist conjecture thus becomes: Given a theory rich in corroborated novel predictive success, theory-parts that have both gained elucidation from independently successful theories and subsequently shown expanded predictive power (very probably) yield approximately true assertions. Clear examples of this strong form of successful elucidation include the derivation of the transversal character and transmission of light waves from Maxwell's theory, sub-atomic elucidations of classical atomic physics and chemistry, molecular elucidations of assertions in genetics, physiology, neurology, including claims previously assumed in medical therapeutic theories, evolutionary assumptions, among numerous other cases. Still, elucidation does not make a part fool-proof. Fortunately, other ways to strengthen the credibility of theory-parts are also available, and they too must be brought into the realist strategy.

[4] I am referring particularly to Worrall (1989) and Leplin (1997).

[5] Whewell (1968) and Darwin (1859).

3 Other Realist Strategies

Explicitly assessing theory-parts as yielding either very probably "true" or very probably "false" assertions is notoriously difficult. Still, identifications of "credible" and "not credible" theory-parts are routinely made in science. Typically, they take time to unveil, but they are there. The standard (albeit frustrating) approach to identification looks for a criterion grounded in either general constraints from just deductive logic and rudimentary induction or general metaphysical views about the relation between theory-parts and the world. Alternatively, realists can try to seize on the kinds of explicit reasons scientist advance when they take specific theory-parts seriously. History provides resources to this end, particularly records of modes of inter-theory comparison and forceful probing regularly applied to theories in mature scientific disciplines. Such probing occurs in general along several overlapping fronts of critical scrutiny, each a source of confirmational import, in addition of the enhanced mode of elucidation already commented on.

To make the matter more vivid I will focus on a case dear to both realists and antirealists: the theoretical development of optics in the nineteenth century. Numerous unreliable and reliable components became gradually disclosed through critical scrutiny along various investigative fronts, each an independent source of confirmational import, five in particular. The first two fronts follow Whewell's emphasis on probing theoretical proposals. In the controversies that led to the great theories (his choicest example was the rise of Newton's mechanics and theory of gravitation), Whewell notes that the conceptions involved were "turned in all directions, examined on all sides; the strength and the weakness of the maxims which men apply to them are fully tested; the light of the brightest minds is diffused to other minds." (1858/1968, p. 109). Application of these insights to theory-parts gives F1 and F2:

(F1) Hostile Probing: Corpuscularians reacted to wave theory by laboring to show that Fresnel's basic tenets were wrong, most famously in the episode leading to the experimental demonstration of the so-called "Poisson Spot" – a prediction that Simeon-Denis Poisson and other corpuscularians thought would ruin Fresnel's theory. To their surprise it crowned it. Conducted primarily by opponents in reaction to a theory's early success, this kind of probing challenges and tries to do without the central tenets of a theory.

(F2) Probing of Auxiliary Assumptions: Thomas Young's experimental demonstration of the phenomenon of double-slit interference in 1801 forced corpuscularians into elaborate auxiliary hypotheses to account for the phenomenon. The latter failed to satisfy and promptly led to the effective collapse of the particle-optics camp. Supporters of a theory typically embark on this line of probing upon encountering difficulties in applying a theory. Success is not guaranteed, as the noted corpuscularian efforts illustrate.

These two fronts work best together, their combined action generally producing two kinds of relevant results. At one extreme, action along these fronts exposes theory components on which numerous problematic cases converge,

> which are thus highlighted as very probably false, as in the example given in F2. At the other extreme, action exposes theoretical parts that seemingly cannot be removed without bringing the theory to stagnation, thus giving a strong indication that those parts are indispensable and very probably truth-worthy, as in the example given for F1.

Front (3) corresponds to the resource of elucidation discussed in the previous section.

(F3) External explanation (strong version of elucidation) of assumptions in a theory: This strategy, discussed in the previous section, seeks to account for initially unsupported or weakly supported claims in terms of independently successful theories. An optical instance in point is the explanation of the refractive index of the various media in terms of their atomic structure. As already said, this front's weakness lies in the comparative unreliability of explanations (recall the examples considered in Sect. 2). On the other hand, elucidation of a theory-part in theory Ta from an initially remote theory Tb brings support less likely to be mortgaged to the conceptual underpinnings of Ta.

Two further diachronic DAC strategies, frequent in scientific theorizing, need to be included as well. One (Front 4) seizes on the articulation of adequacy conditions for theory revision; the other (Front 5) seizes on explanation of the success of superseded theories from the vantage point of current theory. Contrary to appearances, none of these DAC strategies amount to post-hoc maneuvering, but they have other weaknesses.

(F4) Efforts to identify adequacy conditions for future theories: These occur when a proposal faces persistent difficulties and scientists begin to look for alternatives. Parts of the current theory thought to be particularly trustworthy are selected as adequacy conditions for successor theories (correspondence rules, limiting cases, etc.), as when wave theorists took the laws of color separation from the corpuscular theory of light as adequacy conditions. This strategy has sometimes classified chaff as wheat (recall the confidence with which mainstream physicists took the light-ether to be beyond reasonable doubt until the 1920s). The problem is serious because it shows that this front is easily marred by metaphysical intrusion. On the whole, however, this front has a good track record of picking posits subsequently accepted as approximately correct in a stable way (the character of Fresnel's light waves, the kinetic theory of matter, classical chemical structures, conservation principles, and so forth).

(F5) Explanation of the Success of Superseded Theories, as in the way Maxwell's theory provided an informative electromagnetic explanation of Fresnel's postulates regarding the undulating and transversal character of light. Elucidations of this kind typically go hand in hand with theory replacement that deepens specific parts of an earlier theory. They do this by providing a causal and/or structural explanation for tenets of the replaced theory. Often this fifth line of identification also provides hints (in some cases even stable determinations) of the scope and accuracy of retained theory-parts relative to present theory. Contrary to first appearances, none of the features appealed to here involves "vicious post-hoc maneuvers." The maneuvering involved is “virtuous” in that seeking to explain the success of a theory (T0) from the vantage point of a successor (T1) generally contributes epistemic gains along two complementary lines: (A) specification of *divergences* between T0 and T1 that point to *novel predictions* from T1 (relative to T0), a contribution which then contributes to the epistemic evaluation of T1 along F1 and F2. (B) Specification of regimes or regions in the logical space of T0 over which theoretical descriptions drawn from T0 are correct (from the vantage point of T1). In particular, this front contributes hints (in some cases even determinations) of the scope and accuracy of retained theory-parts relative to present theory. These considerations seem especially relevant to naturalist realist projects that admit that all interesting “whole theories” are probably false. Front 5 helps naturalist realism by enhancing the coherence of taking a realist stance about parts of a discarded theory from the perspective of a successor theory also expected to be found faulty in some respect.

Notice that the proposed realist contributions of fronts 4 and 5 are neither guaranteed nor trivial. In the history of science this embedding of theoretical posits and structures from early theories began in earnest only when novel predictive power gained recognition as an epistemic virtue. The transitions from one Ptolemaic theory to another generally display no common theoretical parts (the cycles, epicycles and so on involved are usually very different), the shared descriptions limited to the observable level. Descartes’ Vortices theory displays likewise agreement with Newton’s theory.

All the listed fronts fail sometimes, and do so in ways that compromise entire domains of theorizing, which renders the fronts poor as realist markers. Nevertheless, the historical record seems to support this much: Jointly, the five fronts have singled out, in diachronic fashion, components worthy of realist commitment. This is so not just in the case of optical theories but also in many other theories with similarly clear credentials in terms of corroborated novel predictions. Arguably, it was combined filtering of the suggested types that raised the scientific trustworthiness of such claims as that, in several key respects, light is as Fresnel said and atoms are as classical physics portrayed them; material transformations are to an impressive extent as pre-quantum chemistry said; the evolution of many species is largely as Darwin’s original proposal stated. A preliminary glance at other scientific episodes suggests that the five fronts generalize well. The general claim, then, is that application of the streams of probing to once successful but now superseded theories of the last 250 years identifies arrays of thickly textured (if usually domain-restricted and coarse-grained) descriptions and narratives about several aspects of the world (notably about underpinning structures, microscopic entities, and origins and evolutionary histories).

Drawing from scientific practice, therefore, I suggest that the encountered epistemic shortcomings of explanatory elucidation can be compensated for with additional epistemic support for theory-parts from the other confirmational fronts just highlighted.

4 Theory-Parts for Realists

Here is a tentative proposal, based on the previous reflections. Building on scientific practice and Lakatosian analyses of theory testing along fronts akin to F1, F2 and F3, particularly (Balashov 1994), together with analyses of theoretical deepening by Kitcher (1984) and Thagard (2007), I suggest a criterion that seemingly characterizes theory-components likely to be retained from theories enjoying novel predictive success – components that realists could then identify as true either directly by induction or though inference to the best explanation.

Criterion DAC: As a theory T is applied to diverse situations and the diachronic fronts previously outlined act on T, significant theory-parts gain support of the following sorts[6]:

1. **Refutational DAC** is a criterion of falsehood-worthy: A given theory-part will reveal itself as "doubtful" if multiple pieces of recalcitrant data converge inferentially in that specific part, and saving it is consistently accompanied by degeneration of the whole system, as measured by current epistemological criteria. In many cases a part that receives negative sanction remains in *generalized form* in the successor theory. Examples from classical physics include the principle of strict mass conservation and the absolute simultaneity relation, both of which appear in Special Relativity as approximately correct claims for negligible values of v/c. The important point is that their original versions as exact relations got refuted.
2. **Soft-DAC** articulates a moderate level of positive warrant: A theory-part P will reveal itself as "probably approximately-true" if it is *either* (a) implicated in the theory's empirical success to the point that removing or changing P has consistently led to empirical degeneration; or (b) P has gained elucidation (deepening) from some independent theory rich in corroborated novel predictions.

   Components (2a) and (2b) are diachronic counterparts to KLP's synchronic strategy for identifying theory-parts worthy of realist commitment. The prolonged retention enjoyed by the ether posit warns against (2a). Against (2b) stand cases like Fresnel's successful labors to "substantiate" the geometrical approach to optics (firmly in place before his time) by embedding it in his ether theory, a feat attempted after the latter had guided his theorizing to the successful prediction that made wave optics "irresistible." Soft-DAC is not the strongest realist stance encouraged by the previous sections. Still, soft-DAC identifies specific theory-parts. At various levels (fundamental as well as intermediate), and acting over short time spans, F1 and F2 generally single out some specific theory-parts as components consistently implicated in the successful predictions of a theory, and also other specific components as parts multiply implicated

[6] Typically, these parts correspond to applications of T grounded in models subjected to abstraction, domain-restriction, and coarse-graining (partial-models of T). See Cordero (2013).

in its problematic applications. Both these fronts are insensitive to the kind of metaphysical intrusion that mars the KLP approach, but their historic record is better than the latter's. F3 fails to filter out some "false positives," but it often counteracts metaphysical intrusion, and its track record is also significantly above average. F4 takes stock of the situation, including the yields of F1, F2 and F3. As noted, F5 helps to recognize truth and approximate truth in earlier theories, and also moves accreditation of the successor theory along F1 and F2. Both F4 and F5 display superior track records. It seemingly follows, therefore, that soft-DAC, although only moderately warranted, does go well beyond the very general, vague, and in common "consolation realism" found in the literature ("realism about whatever is actually responsible of the empirical success"). Admittedly, however, components (2a) and (2b) do not do very greatly on their own. On the other hand, their *combined* strength is something to reckon with. That strength fuels a more demanding version of DAC.

3. **Hard-DAC** provides stronger positive warrant than soft-DAC: A theory-part P will reveal itself as "very probably approximately-true" if both conditions (a) *and* (b) above obtain. Conjoining (2a) and (2b) makes, I suggest, for a condition strong enough to give us the position DAC realists seek to provide. Hard-DAC identifies as very probably true only those specific claims that satisfy both of the conditions listed. Its historical record seems excellent.

   Realist explanations of various strengths follow accordingly. Refutational-DAC helps remove commitment to dubious constructs but does not yet warrant a realist position. A moderate realist thesis follows from soft-DAC, while hard-DAC makes for a stronger (although still fallible) realist position. None of the proposed criteria resorts to retrospective projection. Cases like those of the caloric theory, phlogiston, and the nineteenth century ether of light seemingly support the proposed three-fold criterion. More common historical cases appear to do likewise – e.g. the transitions involving Newtonian mechanics, Newtonian gravitation, pre-quantum chemistry, quantum mechanics, Einstein's photon theory, and evolutionary histories in biology, among others.

   How trustworthy is hard-DAC realism? While everything in the resulting scientific picture remains defeasible, it seems (paraphrasing Whewell on consilience[7])that no example can be pointed out, in the whole history of science, in which filtering of the noted kinds together with explanatory elucidation has given testimony in favor of a theory-component subsequently discovered to be false. Of course, for any proposed criterion to be inductively compelling, the history of science would have to endorse the criterion "for the most part." That strongly seems to be the case for the mature period of the scientific disciplines

[7] In Whewell (1847, pp. 67–68) he says: "No example can be pointed out, in the whole history of science, so far as I am aware, in which this Consilience of Inductions has given testimony in favour of an hypothesis afterwards discovered to be false. … [W]hen the hypothesis, of itself and without adjustment for the purpose, gives us the rule and reason of a class of facts not contemplated in its construction, we have a criterion of reality, which has never yet been produced in favour of falsehood".

mentioned in this paper, i.e. from about the 1830s, when independent support and novel prediction gained strong recognition as epistemic values. Subsequent scrutiny will tell.

Acknowledgement I wish to thank an anonymous referee for valuable comments.

References

Balashov, Y. (1994). Duhem, Quine, and the multiplicity of scientific tests. *Philosophy of Science, 61*, 608–628.

Bohm, D. (1957). *Causality and chance in modern physics*. London: Routledge & Kegan Paul.

Carrier, M. (2004). Experimental success and the revelation of reality. In M. Carrier et al. (Eds.), *Knowledge and the world* (pp. 137–161). New York: Springer.

Chakravartty, A. (2003). The structuralist conception of objects. *Philosophy of Science, 70*, 867–878.

Chakravartty, A. (2007). *A metaphysics for realism: Knowing the unobservable*. Cambridge: Cambridge University Press.

Chang, H. (2003). Preservative realism and its discontents: Revisiting caloric. *Philosophy of Science, 70*, 902–912.

Cordero, A. (2011a). Rejected posits, realism, and the history of science. In *The European Philosophy of Science Association Proceedings* (Vol. 1, pp. 23–32). Dordrecht: Springer.

Cordero, A. (2011b). Scientific realism and the *divide et impera strategy*: The ether saga revisited. *Philosophy of Science, 78*, 1120–1130.

Cordero, A. (2013). Conversations across Meaning Variance. *Science & Education 22*, 1305–1313.

Darwin, C. (1859) On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray.

Davis, A. E. L. (2003). Conversations across Meaning Variance. *Science & Education 22*, 1305–1313.

Kepler, J. (1620/1995). *Epitome of Copernican astronomy,* Book IV (trans: Charles G. Wallis). Amherst: Prometheus Books.

Kitcher, P. (1984). 1953 and all that: A tale of two sciences. *Philosophical Review, 93*, 335–373.

Kitcher, P. (1993). *The advancement of science*. Oxford: Oxford University Press.

Laudan, L. (1984). A confutation of convergent realism. In J. Leplin (Ed.), *Scientific realism* (pp. 218–249). Berkeley: University of California Press.

Leplin, J. (1997). *A novel defense of scientific realism*. Oxford: Oxford University Press.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. London/New York: Routledge.

Sulloway, F. (1992). *Freud, biologist of the mind: Beyond the psychoanalytic legend*. Cambridge, MA: Harvard University Press.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science, 74*, 28–47.

Whewell, W. (1847). *The philosophy of the inductive sciences* (Section III – Tests of hypothesis, Vol. II). London: John W. Parker.

Whewell, W. (1858/1968). Chapter III, Induction and scientific method. In E. Robert (Ed.), *William Whewell's theory of scientific method* (pp. 103–264). Pittsburgh: University of Pittsburgh Press.

Worrall, J. (1989). Fresnel, Poisson and the white spot: The role of successful prediction in the acceptance of scientific theories. In G. Gooding et al. (Eds.), *The uses of experiment* (pp. 135–157). Cambridge: Cambridge University Press.

Natural Kinds and Concept Eliminativism

Samuli Pöyhönen

Abstract Recently in the philosophy of psychology it has been suggested that several putative phenomena such as emotions, memory, or concepts are not genuine natural kinds and should therefore be eliminated from the vocabulary of scientific psychology. In this paper I examine the perhaps most well known case of scientific eliminativism, Edouard Machery's concept eliminativism. I argue that the split-lump-eliminate scheme of conceptual change underlying Machery's eliminativist proposal assumes a simplistic view of the functioning of scientific concepts. Conceiving of scientific concepts as natural kind terms is an important reason for the impasse between Machery and anti-eliminativists, as both sides allude to properties of natural kinds in their contradicting arguments. As a solution I propose that, in order to develop a more satisfactory theory of conceptual change in science, one needs to distinguish between three different types of scientific concepts, hitherto conflated under the loaded notion of natural kind.

1 Introduction

Eliminativism has a venerable history in the philosophy of mind and the philosophy of psychology, and the arguments for abandoning independent mental substances and properties from our ontological catalogue have played an important role in the development of both philosophical as well as scientific thinking about the mind. In the latter part of the twentieth century, eliminativist arguments were often directed at mental states posited by common sense psychology, and in the 1980s they received broad attention in the debates over eliminative materialism. Recently a new mechanistic variant of eliminativism has emerged. Based on a model of

S. Pöyhönen (✉)
TINT/Social and Moral Philosophy, University of Helsinki, Helsinki, Finland
e-mail: samuli.poyhonen@helsinki.fi

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_14,

conceptual change that I call *the split-lump-eliminate scheme* (SLE scheme), it has been suggested that familiar notions such as EMOTION, MEMORY, or CONCEPT do not correspond to genuine natural kinds, and should therefore be eliminated from scientific vocabulary (Griffiths 1997; Machery 2005, 2009; Piccinini and Scott 2006). Edouard Machery's *concept eliminativism* is perhaps the most hotly debated example of these recent eliminativist projects. According to Machery's heterogeneity hypothesis, the human capacity for conceptual thought is supported by at least three different kinds of representations and processes, and thus CONCEPT is not a natural kind. However, a large group of philosophers and psychologists alike have resisted Machery's eliminativist conclusion (see the peer commentary on Machery 2010). It has been a common reaction to Machery's position to argue that despite the differences between different kinds of concepts, the notion has an important theoretical role in psychology, and thus cannot be abandoned.

Machery presents a strong case for the claim that CONCEPT is not a useful notion for describing open explananda in psychological research on human conceptual abilities, i.e., psychological phenomena whose properties are still at least partly unknown and that are under ongoing inquiry. However, this would be a sufficient reason for concept eliminativism only if referring to explananda was the only epistemic function for scientific concepts.[1] This is where I part ways with Machery and side with the defenders of concepts. I contend that functionally identified kinds sustained by abstractly characterized causal mechanisms often play an epistemically important explanatory role in the sciences. Hence, if CONCEPT turns out to be such a functional kind, its heterogeneity alone is not a sufficient reason for elimination. Machery's eliminativist inference is therefore premature, and his heterogeneity hypothesis must be qualified for it to be sound.

My diagnosis of the conflict between Machery and the anti-eliminativists about concepts is that the disagreement can be traced to a simplistic picture of the functioning of scientific concepts, *the natural kinds model*. To reveal the inadequacy of this approach shared by both sides of the debate, I introduce the SLE scheme and its problematic application to CONCEPT in Sects. 2 and 3. I then examine the notion of natural kind underlying the model. I show that, unlike what its proponents suggest, in most cases the SLE scheme does not provide unambiguous recommendations for conceptual change in science, and fails as a normative foundation for concept eliminativism. In Sect. 6, I offer my positive contribution. I introduce a distinction between three types of scientific concepts that hitherto have been conflated under the notion of natural kind. My threefold division of kinds classifies scientific

[1] In this paper, the term 'concept' appears in two different meanings. 'Concept' in the psychological sense refers to a putative cognitive structure of an individual. 'Scientific concepts', on the other hand, are things featured in scientific theories and employed collaboratively by scientists in their research practices. Although there certainly are continuities between these two uses of the term, it is important to clearly distinguish between them.

concepts according to their epistemic role.[2] Roughly, *investigative kinds* are vehicles for representing targets of ongoing empirical research, *instrument kinds* function as explanantia, and non-mechanistic *framework kinds* are tools for coordination between different research perspectives on complex targets of scientific inquiry. My distinction draws attention to an important dimension of conceptual change overlooked by the SLE scheme of scientific eliminativists. I suggest that changes in the inferential potential of a concept constitute an aspect of conceptual change not reducible to alignment of concepts with causal structures in reality.

2 The Split-Lump-Eliminate Scheme

The SLE scheme underlying Machery's concept eliminativism builds on the idea that scientific concepts should refer to natural kinds. I call this approach the *natural kinds model of scientific concepts.* The theory of natural kinds employed by Machery and other scientific eliminativists is an interpretation of Richard Boyd's homeostatic property cluster theory (HPC), according to which natural kind concepts should be aligned with causal mechanisms in reality. According to Boyd's (1991, 1999) theory, a natural kind is characterized by

(α) a cluster of typical properties that is supported by
(β) a homeostatic mechanism that brings about their co-occurrence.

The SLE scheme of conceptual change, which builds on this foundation, is based on three operations: If a concept refers to several different mechanisms, one should *split* it so that each mechanism gets its own corresponding concept. On the other hand, a concept should capture the maximal class of phenomena sustained by the same mechanism. Therefore, if we can find the same mechanism behind a group of phenomena that were previously considered as separate, we should *lump* them under the same concept. And thirdly, were it to turn out that there is no well-defined mechanism corresponding to a concept, we should *eliminate* this notion from scientific usage. In sum, the core idea underlying these three operations is that there should be a one-to-one correspondence between scientific concepts and mechanisms in reality (Griffiths 1997, 2004; Machery 2009; Craver 2009).

In comparison to probably the most well-known account of conceptual change in the philosophy of psychology, eliminative materialism, the SLE scheme is an advancement in several respects. First, it offers a more fine-grained picture of conceptual revision by not only focusing on elimination but by also including cases of unification and non-eliminative conceptual refinement as species of conceptual

[2] See Brigandt (2010) for a somewhat similar picture of the conceptual dynamics of science. In attempting to account for the rationality of conceptual change, Brigandt emphasizes inferential role and epistemic goal as important semantic dimensions of scientific concepts.

change. Second, the model does not rely on semantic intuitions about reference as a basis for conceptual change, but instead draws on the widely accepted realist judgment that our scientific classifications ought to be aligned with the causal structures of reality. It thus conceives of the conceptual dynamics of scientific psychology as being continuous with those of other scientific fields, whereas the domain of eliminative materialism is limited to folk psychological predicates only. Moreover, as the work of scientific eliminativists suggests, perhaps the most convincing evidence in favor of the SLE scheme comes from its ability to account for several recent episodes of theoretical development in the human sciences (cf. Griffiths 1997; Craver 2004; Wilson et al. 2007; Machery 2009).

3 Concept Eliminativism and its Discontents

The instance of scientific eliminativism that has recently raised the most debate is Edouard Machery's concept eliminativism (2005, 2009, 2010). In *Doing Without Concepts*, Machery defines concepts in psychology in the following way:

> A concept of x is a body of knowledge about x that is stored in long-term memory and that is used by default in the processes underlying most, if not all, higher cognitive competences when these processes result in judgments about x (Machery 2009, p. 12).

Based on empirical research in the cognitive sciences, Machery (2009, p. 4) then formulates his heterogeneity hypothesis:

1. The best available evidence suggests that for each category an individual typically has several concepts.
2. Co-referential concepts have very few properties in common. They belong to very heterogeneous kinds of concept.
3. Evidence strongly suggests that prototypes, exemplars, and theories are among these heterogeneous kinds of concept.
4. Prototypes, exemplars, and theories are typically used in distinct cognitive processes.
5. The notion of concept ought to be eliminated from the theoretical vocabulary of psychology.

Machery reviews plenty of empirical evidence for each of the tenets 1–4, and despite dealing with contestable issues, they have raised relatively little controversy. Not so for tenet 5, Machery's normative conclusion and the main result of his book. It appears that Machery regards the last tenet as an implication of the conjunction of tenets 1–4 together with the principles described above as the SLE scheme. According to Machery, evidence suggests that scientifically interesting generalizations about concepts are actually sustained by mechanisms corresponding to the different subkinds, and because these mechanisms are sufficiently distinct, there is no well-defined mechanism underlying CONCEPT as such. Based on the SLE scheme, Machery then concludes that the notion of concept should be eliminated from scientific psychology, and replaced with lower-level notions referring to prototypes, exemplars, and theory-based concepts.

Perhaps the most common challenge to Machery's eliminativist conclusion has been to emphasize the indispensable theoretical role that the notion of concept plays in psychological research. Machery's critics have argued that only CONCEPT captures a set of questions and generalizations that have to do with the human capacity for conceptual thought in general. Abandoning the notion would therefore deemphasize this set and hinder scientific progress because there would be no notion to integrate results from research on subkinds of concepts (Couchman et al. 2010; Edwards 2010; Hampton 2010). A problem with many of these replies is that while they draw on psychologists' intuitions about the epistemic role of the notion of concept, they have not often been based on systematic theories of the functioning of scientific kind terms.

However, Richard Samuels and Michael Ferreira (Samuels and Ferreira 2010) have replied to Machery on his own ground. They argue, in contrast to Machery's claim, that there are good reasons to accept CONCEPT as an HPC natural kind. First, there is a reliably occurring property cluster associated with the kind:

1. Concepts consist in bodies of information, and are
2. Stored in long-term memory,
3. Promiscuous (the same information is employed by several higher cognitive abilities),
4. Internally connected, and
5. Internally coherent.

In his defense of concept pluralism, Daniel Weiskopf (2009) has introduced some further shared properties of concepts:

6. Concepts are sensitive to logical form,
7. They combine productively and systematically, and
8. Are acquired by employing similar cognitive processes.

The argument thus goes that CONCEPT has a proprietary cluster (α) of projectible properties. Secondly, Samuels and Ferreira (2010, p. 222) suggest that this property cluster is sustained by a functionally identifiable causal process: the cognitive mechanism (β) corresponding to the cluster above is closely related to processes behind long-term memory, and their relations to other higher cognitive processes. Together these considerations suggest that CONCEPT would qualify as an HPC natural kind. Henceforth, I call this way of defending concepts the *anti-eliminativist position*.

Importantly, Samuels and Ferreira largely agree with Machery on the empirical facts about concepts, but deny his normative conclusion. The sticking point appears to be the correct level of description of our conceptual abilities. To shed light on the disagreement, in the next section I examine the role and motivation of natural kind taxonomies in science.

4 Why Worry About Natural Kinds?

The notion ‘natural kind’ is discussed in several areas of philosophy, and in different contexts it serves slightly different conceptual aims. In the philosophy of language it has played a central role in arguments against descriptivism, and in metaphysics, the concept features in discussions concerning laws of nature, natural necessity, and essentialism (cf. Bird and Tobin 2008). The discussions on natural kinds in the philosophy of science constitute the third, partly independent strand of the tradition of natural kinds (Hacking 1991). Within this epistemology-oriented approach, questions of natural kinds concern primarily scientific concept formation (Reydon 2009). It is a common intuition that only when our concepts correspond to natural kinds, do they succeed in referring to genuine phenomena in reality, and can be reliably employed in the epistemic practices of prediction, explanation, and manipulation of these phenomena. Moreover, it is now agreed by many that since the epistemic aims of the human sciences are similar to those of natural sciences, the categories in the human sciences should also conform to the natural kinds model (Sterelny 1990, Chap. 3; Boyd 1991). Thus, from the perspective of the philosophy of science, the reason to worry about the natural kindhood of concepts is primarily in order to maximize epistemic power and reliability.

However, even within the philosophy of science, there are multiple competing conceptions of what being a natural kind amounts to. Often the notion is used without explicating what is exactly meant by something being a natural kind, but a review of the current literature suggests the following list of criteria commonly attributed to natural kinds in the special sciences (cf. Hacking 1991; Boyd 1999; Murphy 2006, Chap. 9; Bird and Tobin 2008; Samuels 2009):

{NK}

1. Induction justification: Natural kinds should license inductive inferences.
2. Causal grounding: Natural kind concepts should track the causal structure of reality. The unity of a kind is causal, not conceptual.
3. Non-analyticity: Members of a natural kind share a large number of (logically unrelated) non-trivial properties in addition to the ones that are used to identify the kind.
4. Semantic open-endedness: The semantics of natural kind terms is such that it makes sense to attempt to refine their meaning through empirical inquiry.
5. Lawfulness: Natural kinds are referred to in laws of nature.
6. Essentialism: Natural kinds have essences constituted by their intrinsic properties.
7. Uniqueness: There is a unique best taxonomy of reality in terms of natural kinds that represents nature as it is.

The list is not meant as an exhaustive characterization of the properties of natural kinds – alternative suggestions abound. Moreover, most theories of natural kinds only subscribe to some of the criteria. In fact, I suggest that different combinations of the criteria can be used to isolate somewhat independent dimensions in the meaning of 'natural kind,' which I employ in Sect. 6 in distinguishing between different epistemic roles that kind concepts play in scientific research practices.

Assessing HPC theory in the light of {NK} reveals its liberal nature. The definition of HPC kinds offered in Sect. 2 suggests that they obviously satisfy criteria 1 and 2: an HPC kind consists of a reliably occurring cluster of projectible properties allowing for reliable extrapolation, and its epistemic reliability follows from the kind being anchored in existing causal structures. However, HPC theory is not committed to the whole group of the remaining criteria: The concept of mechanism has a central role in the theory primarily in order to avoid employing the problematic notion of law of nature. Moreover, several proponents of the theory have emphasized that mechanisms underlying kinds need not consist only of intrinsic properties of kind members (Boyd 1991; Griffiths 1997; Murphy 2006). HPC theory is thus not committed to properties 5 or 6.

Moreover, as a recent argument by Carl Craver (2009) shows, HPC classifications do not satisfy the uniqueness criterion (7). In brief, drawing on recent research on the notion of causal mechanism, Craver observes that while mechanisms are constituted by real causal structures, decisions regarding the correct level of mechanistic description and demarcation of the boundaries of mechanisms require considerations of explanatory relevance. This is because descriptions of mechanisms are explanatory devices – mechanisms are identified in order to explain properties of an explanandum phenomenon. Although "kinds are where the mechanisms are," mechanism individuation in turn requires prior fixing of explananda.

Hence, mechanisms sustaining HPC kinds can be identified at various levels of abstraction, depending on the epistemic aims of the research perspective.[3] Here HPC theory seems to capture a genuine aspect of scientific practice: examples of classificatory pluralism abound in the life sciences. As Robert Richardson (2008) and Mark Couch (2009) have observed, "multiply realized" concepts such as EYE or ENZYME capture sufficiently homogeneous units for certain epistemic purposes, whereas from other perspectives they appear as heterogeneous kinds. Such higher-level categories are usually functionally individuated, and their corresponding mechanisms are abstractly specified causal structures. In sum, once the often-unanalyzed notion of mechanism is spelled out properly, it turns out that *causally sustained functional kinds qualify as HPC kinds*. Therefore, it appears that criteria 3 and 4 are not necessary conditions for HPC kinds: If the abstract mechanism exhausting the unity of the kind (e.g., ENZYME, TURING MACHINE) is already known, it does not appear useful to further examine the internal makeup of the members of the kind in order to learn more about the cluster of kind-properties ($\sim$4). Moreover, in cases in which the relationship between the mechanism and the

[3] Boyd (2010) himself recognizes this conventionalist aspect of scientific classification, when he states that there are no kinds that are natural *simpliciter*, but instead kinds are natural with respect to the inferential architectures of particular disciplinary matrices.

corresponding kind properties is transparent, it is often questionable whether it can meaningfully be said that the kind is characterized by a large group of properties not accounted by the definition (i.e., the mechanism description) of the kind.[4]

5 Inadequacy of the SLE Scheme

This analysis of the commitments of HPC theory points to a serious shortcoming in the SLE scheme: In situations where our classifications are outright incorrect and do not correspond to any well-defined causal mechanisms (e.g., sublunary objects, phlogiston, phrenology, drapetomania), the model rightly suggests elimination. However, in most cases discussed by scientific eliminativists, this is not the case. Instead, decisions of splitting and lumping are often done between causally sustained classifications at different levels of abstraction. This is also the case with concept eliminativism: both the lower-level categories (prototype, exemplar, and theory-based concepts) as well as the higher-level notion of concept facing elimination can be understood as mechanistic HPC kinds. In such cases, none of the operations of the SLE scheme apply.

It appears that in order to save the eliminativist conclusion, supporters of the SLE scheme must adhere to a stricter notion of natural kind. Perhaps the intuitively most obvious option would be to require that members of a natural kind must share the same internal structure. However, this option is not open in the domain of psychology. Cognitive kinds in general are supported by abstractly characterized mechanisms: they are implemented in plastic neural structures, and therefore implementation-level differences between their instances are unavoidable.

In his response to concept pluralists, Machery (2009, pp. 243–245) has adopted a different strategy. He claims that there has to be further empirically discoverable generalizations to be made about natural kinds, and thus in treating functional kinds as natural, concept pluralists misconceive the nature of natural kinds. Machery thus simply assumes that natural kinds should satisfy criteria 3 and 4. This judgment appears to stem from the fact that heterogeneous categories like CONCEPT are not plausible *explananda*, as empirical research conducted by using such notions would result in disjunctive theories and explanations. However, this move begs the question against pluralists because, for them, the ability to ground reliable theoretical inferences and explanations – to function as an *explanans* – is sufficient for CONCEPT's kindhood (cf. Weiskopf 2009). As suggested in the previous section, functionally individuated HPC kinds can serve this epistemic purpose: they are by definition characterized by reliably occurring property clusters sustained by abstract causal mechanisms. Therefore, as long as the generalizations made by employing a

[4] Peirce (1903) observed this tension already in Mill's account of kinds: Mill (1891) requires that a small group of properties must not account for the rest of a real kind's properties but, on the other hand, the aim of scientific research to find law-like relationships between the properties of kinds appears to undermine their independence.

concept concern this property cluster, the implementation-level differences between instances of the kind can safely be ignored. Functionally individuated HPC kinds satisfying only criteria 1 and 2 can thus be treated as inferential tools that "black-box" the non-pertinent implementation-level differences between instances of the kind.

Hence, Machery's eliminativist conclusion appears to be blocked by a competing account of what natural kindhood amounts to. The competing conceptions emphasize two different but equally important *epistemic roles* that scientific concepts can play. Whereas eliminativism is driven by the idea that natural kind concepts stand for plausible explananda, an anti-eliminativist can emphasize CONCEPT's role as an indispensable explanans in psychological theories. In the following penultimate section of my paper, I suggest that these two distinct epistemic roles should be clearly distinguished, and that trying to prove the primacy of either one of them is a misguided effort, only motivated by the monolithic natural kinds model of scientific concepts. As my positive contribution, I suggest a threefold division between different types of scientific concepts.

6 Splitting the Notion of Natural Kind

My reconstruction of the notions of natural kind used by eliminativists and their opponents suggested that an eliminativist needs to adhere to a notion that includes all of the criteria 1–4 as necessary conditions of natural kinds, whereas anti-eliminativists employ a more liberal notion that only clearly satisfies 1 and 2. In this section I suggest that these clusters of criteria can be used to identify two different types of concepts employed in science, each with their corresponding epistemic niche. Furthermore, I suggest that also non-mechanistic concepts often play an important role in research practices. By grouping types of scientific concepts according to their epistemic roles, we get the following threefold classification:

(A) *Investigative kinds.* Adopting a term from Brigandt (2003) and Griffiths (2004), I call 'investigative kinds' the group of scientific concepts that capture many of the intuitions behind traditional conceptions of natural kinds. Treating a concept as an investigative-kind concept means that in addition to justifying inductive inference, members of the kind are assumed to share yet unknown similarities, and thus we can learn more about them by empirically investigating the properties of their instances. For this reason, investigative kind concepts are vehicles for representing targets of ongoing empirical research, and often stand for explananda in scientific theories. Examples of investigative kinds would include elementary particles and neutron stars, but also psychological explananda such as schizophrenia or confirmation bias.

(B) *Instrument kinds*. Unlike investigative kinds, instrument kind concepts typically function not as explananda but as explanantia: they serve as vehicles for explanation and storage of scientific knowledge. As argued above, despite being functionally identified kinds, they can serve in these epistemic roles because members of the kind share a robust cluster of projectable properties supported by an abstractly specified causal mechanism. However, instrument kinds are not characterized by the same semantic open-endedness as investigative kinds and are thus poor devices for reductive research: there is no reason to assume that their members share non-trivial properties apart from ones governed by the known homeostatic mechanism of the kind. Instead, the epistemic power of instrument kinds like EYE, ENZYME, MARKET, or TURING MACHINE stems from their ability to capture general patterns and abstract mechanisms common to several different targets and domains.

(C) *Framework kinds.* As observed already by Hilary Putnam (1965, p. 379), many central scientific concepts are not defined by their role in a single law or theory, but are law-cluster concepts residing at the intersection of several theories. Putnam's example was ENERGY, but several cases can be found in the human sciences as well: GENE, RATIONALITY, INFORMATION, and REPRESENTATION are examples of important concepts that however have slightly different meanings in different research programs (Griffiths and Stotz 2007; Bermúdez 2005, pp. 9–10). I suggest that despite not being anchored in any specific causal mechanisms, framework kinds often play an important epistemic role. As suggested by Susan Leigh Star in her work on boundary objects, in science we need concepts simultaneously inhabiting several social worlds. They must be malleable enough to adapt to the informational requirements of different disciplines, but still maintain the identity of the target across different sites (Leigh Star and Griesemer 1989). To put the matter in terms of {NK}, framework kinds do not satisfy criteria 1 and 2, but 3 and 4 capture important aspects of their functioning. Open-endedness and indexicality are semantic properties that allow the reference of a concept to be fixed independently of particular descriptions, and framework concepts can thus correspond to targets of research whose mechanisms and best levels of description are still unknown.

This tentative classification of scientific concepts according to their epistemic roles is still coarse, and the details of the proposal need to be worked out. However, the scheme is arguably more useful than the monolithic natural kinds model: It appears that all the three types are manifested in scientific research and correspond to distinct epistemic niches. Moreover, all stand apart from conventional or erroneous classifications. This more refined picture of scientific concepts is also useful for making sense of the debate on concept eliminativism, because it can accommodate both Machery's and anti-eliminativist insights: Machery convincingly shows that CONCEPT does not qualify as an investigative kind, and thus trying to uncover the whole set of projectable properties of concepts would be misguided. On the other hand, in several theoretical contexts in psychology concepts are explanantia rather than explananda (Lombrozo 2011). As argued above, for these purposes it suffices that a concept satisfies the requirements for instrument kinds.

Moreover, my scheme suggests a third possibility. Retaining CONCEPT as an instrument kind requires that it correspond to a well-defined causal mechanism.

The jury is still out on this question, partly due to the fleeting nature of the notion of mechanism, and partly to inconclusive empirical research. Even conceding the eliminativist the judgment that CONCEPT is not a mechanistically grounded kind, the notion could survive as a framework kind coordinating research between several fields investigating higher cognitive abilities (e.g., psychology, social sciences, and AI).

7 Conclusion: The Fate of Concepts

I have argued that not all scientific concepts serve the same epistemic purpose. Working out the consequences of this insight suggests that Machery's eliminativist conclusion does not follow from his heterogeneity hypothesis. However, the genuine insight of Machery's position can be saved by qualifying his argument: Heterogeneity of CONCEPT does not recommend its elimination but it does show that the notion does not pass as an investigative kind, and hence cannot serve the corresponding epistemic role in scientific research practices. Acknowledging this change in the inferential status of the notion can have the same epistemic benefits for psychology as Machery uses to motivate his eliminativist position (cf. Machery 2009, p. 248): being explicit about CONCEPT's status as an instrumental (or framework) kind should discourage useless primacy debates between different theories of concepts and direct attention towards more relevant questions.

My more general aim in this paper has been to highlight an overlooked form of conceptual change in science. In addition to the operations described by the SLE scheme, conceptual change consists also in often-subtle changes in the inferential potential of concepts. The labels 'investigative,' 'instrumental,' and 'framework kind' correspond to such inferential statuses, and keeping track of how scientific concepts move from one concept-type to another is one way of representing such conceptual change. The trajectory of CONCEPT might provide a typical example of the life course of a scientific concept: starting off as an investigative kind, the notion first promotes research on a phenomenon that is considered unitary. However, after the heterogeneity of the processes behind the phenomenon is revealed, the notion might persist as a tool for the storage of higher-level generalizations, or as a more malleable notion coordinating research and communication between different perspectives on the target. During this process of conceptual change, the splitting and lumping operations suggested by the SLE scheme might lead to the emergence of more precise (or more general) mechanistic classifications, but these events need not be accompanied by the elimination of the original kind concepts.

Acknowledgements An earlier version of this paper was presented at the 14th Congress of Logic, Methodology and Philosophy of Science (CLMPS) at Nancy. I thank audiences both at CLMPS and at EPSA11 for their comments. I also thank Petri Ylikoski and Marion Godman for valuable feedback on earlier drafts.

References

Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. London: Routledge.

Bird, A., Tobin, E. (2008). Natural kinds. In Zalta E (Ed.), *The Stanford encyclopedia of philosophy (Summer 2010 Edition)*. http://plato.stanford.edu/archives/sum2010/entries/natural-kinds/.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies, 61*, 127–148.

Boyd, R. (1999). Kinds as the 'workmanship of men'. In J. Nida-Rümelin (Ed.), *Rationalität, realismus, revision* (pp. 52–89). Berlin: Walter de Gruyter.

Boyd, R. (2010). Realism, natural kinds and philosophical methods. In H. Beebee & N. Sabbarton-Leary (Eds.), *The semantics and metaphysics of natural kinds* (pp. 212–234). New York: Routledge.

Brigandt, I. (2003). Species pluralism does not imply species eliminitivism. *Philosophy of Science, 70*, 1305–1316 (Proceedings).

Brigandt, I. (2010). The epistemic goal of a concept: Accounting for the rationality of semantic change and variation. *Synthese, 177*, 19–40.

Couch, M. (2009). Multiple realization in comparative perspective. *Biology and Philosophy, 24*, 505–519.

Couchman, J., Boomer, J., Coutinho, M. V. C., & Smith, J. D. (2010). Carving nature at its joints using a knife called concepts. *Behavioral and Brain Sciences, 33*, 207–208.

Craver, C. (2004). Dissociable realization and kind splitting. *Philosophy of Science, 71*, 960–971.

Craver, C. (2009). Mechanisms and natural kinds. *Philosophical Psychology, 22*, 575–596.

Edwards, K. (2010). Unity amidst heterogeneity in theories of concepts. *Behavioral and Brain Sciences, 33*, 210–211.

Griffiths, P. E. (1997). *What emotions really are*. Chicago: University of Chicago Press.

Griffiths, P. E. (2004). Emotions as normative and natural kinds. *Philosophy of Science, 71*, 901–911.

Griffiths, P. E., & Stotz, K. (2007). Gene. In M. Ruse & D. Hull (Eds.), *Cambridge companion to philosophy of biology* (pp. 85–102). Cambridge: Cambridge University Press.

Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies, 61*, 109–126.

Hampton, J. (2010). Concept talk cannot be avoided. *Behavioral and Brain Sciences, 33*, 212–213.

Leigh Star, S., & Griesemer, J. (1989). Institutional ecology, 'translations' and boundary objects. *Social Studies of Science, 19*, 387–420.

Lombrozo, T. (2011). The campaign for concepts. *Dialogue, 50*, 165–177.

Machery, E. (2005). Concepts are not a natural kind. *Philosophy of Science, 72*, 444–467.

Machery, E. (2009). *Doing without concepts*. New York: Oxford University Press.

Machery, E. (2010). Precis of doing without concepts. *Behavioral and Brain Sciences, 33*, 195–244.

Mill, J. S. (2002[1891]). *A system of logic: Ratiocinative and inductive*. Honolulu: University Press of the Pacific.

Murphy, D. (2006). *Psychiatry in the scientific image*. Cambridge: MIT Press.

Peirce, C. S. (1903). Kind. In James Baldwin (Ed.), *Dictionary of philosophy and psychology.* New York: The Macmillan Company. http://psychclassics.yorku.ca/Baldwin/Dictionary/. Accessed 24 Oct 2012.

Piccinini, G., & Scott, S. (2006). Splitting concepts. *Philosophy of Science, 73*, 390–409.

Putnam, H. (1975[1965]). The analytic and the synthetic. In H. Putnam (Ed.), *Philosophical papers, vol. 2. Mind, language and reality* (pp. 33–69). Cambridge: Cambridge University Press.

Richardson, R. (2008). Autonomy and multiple realization. *Philosophy of Science, 75*, 526–536.

Reydon, T. (2009). How to fix kind membership: A problem for HPC theory and a solution. *Philosophy of Science, 76*, 724–736.

Samuels, R. (2009). Delusions as a natural kind. In M. Broome & L. Bortolotti (Eds.), *Psychiatry as cognitive neuroscience: Philosophical perspectives* (pp. 49–82). Oxford: Oxford University Press.

Samuels, R., & Ferreira, M. (2010). Why don't concepts constitute a natural kind? *Behavioral and Brain Sciences, 33*, 222–223.

Sterelny, K. (1990). *The representational theory of mind*. Oxford: Basil Blackwell.

Weiskopf, D. (2009). The plurality of concepts. *Synthese, 169*, 145–173.

Wilson, R., Barker, M., & Brigandt, I. (2007). When traditional essentialism fails. *Philosophical Topics, 35*, 189–215.

Against the Statistical Account of Special Science Laws

Andreas Hüttemann and Alexander Reutlinger

Abstract John Earman and John T. Roberts advocate a challenging and radical claim regarding the semantics of laws in the special sciences: the statistical account. According to this account, a typical special science law "asserts a certain precisely defined statistical relation among well-defined variables" Earman and Roberts (*Synthese, 118,* 439–478, 1999) and this statistical relation does not require being hedged by ceteris paribus conditions. In this paper, we raise two objections against the attempt to cash out the content of special science generalizations in statistical terms.

1 Introduction

John Earman and John T. Roberts defend a view according to which fundamental physics states laws that are expressed by universal generalizations (which are not qualified by any ceteris paribus condition or proviso). By contrast, the special sciences do not state universal laws but rather statistical generalizations. Since this account of special science 'laws' does not have a name we call it the *statistical account*. According to the statistical account, special science generalizations are to be interpreted either as statements about actual non-strict correlations or as statements that are 'mostly true'. Earman and Roberts claim that the statistical account does not require a qualification by ceteris paribus (henceforth, cp) conditions. As a consequence, cp-conditions are neither needed for the fundamental laws (because

A. Hüttemann (✉)
Department of Philosophy, University of Cologne, Albertus-Magnus-Platz, 50923 Köln, Germany
e-mail: ahuettem@uni-koeln.de

A. Reutlinger
Department of Philosophy, University of Cologne, Richard-Strauss-Str. 2, 50931 Köln, Germany
e-mail: Alexander.Reutlinger@uni-koeln.de

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_15,

they are strict) nor for the special science generalizations. This seems to be a prima facie advantage for the statistical account because the exact meaning of cp-conditions remains controversial.

In this paper, we will leave aside Earman and Roberts's claim about the laws of physics. We focus on special science generalizations and present two objections against the statistical account.

2 Terminology

In order to lay a foundation for our arguments against the statistical account, we review two useful distinctions that are commonly drawn in the recent literature on cp-laws.

Gerhard Schurz (2002) distinguishes *exclusive* and *comparative* cp-laws. Exclusive cp-laws state that systems display a certain behavior provided there are *no* disturbing factors. The disturbing or interfering factors have to be *absent* for the behavior in question to be displayed. Newton's first law may be an example of an exclusive cp-law, as it describes the behavior of a body in the absence of forces. Other cp-laws require that certain (sometimes unspecified) factors remain *constant* as opposed to being absent. An example is the following law: 'if the supply of a commodity increases (decreases), the price decreases (increases)'. The law is a comparative cp-law because it requires that certain factors remain constant (e.g. demand). Cp-laws may be both exclusive and comparative, as, for instance, in the above example. It is not only required that the demand remains constant (which is often explicitly mentioned), it is furthermore tacitly assumed that there are no state interventions, natural catastrophes, wars etc. Strictly speaking, exclusive cp-laws can be reconstructed as special cases of comparative cp-laws with the relevant variables set to the value 0. However, exclusive cp-laws play a major role in the context of idealizations and special treatments of this case have been suggested. We follow the literature in distinguishing exclusive from (other) comparative cp-laws.

Earman and Roberts introduce a second helpful distinction (that is independent of the first distinction). The distinction of lazy and non-lazy cp-conditions.[1] They argue that a cp-clause is dispensable if all exceptions to the law (and other conditions that have to obtain in order for the generalization to be true) *can* be listed and it is merely a matter of convenience and the result of "laziness" that the conditions are not listed explicitly. Earman and Roberts refer to such a finite list as a "lazy" cp-clause. According to Earman and Roberts, only "non-lazy" cp-clauses are proper cp-clauses: a proper cp-clause is an open clause of which we do not know how to complete it. For what follows we will distinguish two senses of non-lazy, both of which can be found in Earman and Roberts (1999).

[1] See also Earman et al. (2002, p. 283f). Schurz uses the terminology of definite versus indefinite cp-conditions for the same distinction (Schurz 2002, Sect. 3).

i. *Non-Lazy*$_1$: According to the first reading of non-laziness, the list of exceptions and conditions is open-ended and thus cannot be completed (Earman and Roberts 1999, p. 439, 441, 444, 467).
ii. *Non-Lazy*$_2$: According to the second reading, disturbing factors that might need to be taken into account in order to complete the cp-clause are outside the conceptual and methodological resources of the special science in question and, thus, cannot be captured. (Earman and Roberts 1999, p. 462f).

Cp-laws or cp-clauses that are non-lazy$_2$ need not be non-lazy$_1$. Even if the relevant conditions cannot be stated in the vocabulary of the special science, there might be a finite list if we allow for further conceptual resources (of, for instance, the physical sciences).

3 Lange's Dilemma

The problems concerning cp-laws are usually introduced by way of a dilemma, according to which law statements of the special sciences are either false empirical or trivially true statements. Many laws, such as Galileo's law are false if the law is read as a strict (universal) generalization. The claim 'whenever a body falls, it falls according to the equation: $s = ½ gt^2$, is false, because in water and other media the equation does not correctly describe the behavior of the bodies in question. Similarly the claim 'if the supply of a commodity increases (decreases), the price decreases (increases)' is false if read as a strict generalization, because there may be state interventions and other factors which lead to counter-instances to the strict generalization. This is the *first horn* of the dilemma ("falsity").

If, on the other hand, the law is hedged by a cp-clause, then Galileo's law becomes 'whenever a body falls (freely), it falls according to the equation: $s = ½ gt^2$, *unless some interfering factor intervenes*'. This claim appears to be trivially true, at least if the notion of an interfering factor is not further specified. If what is meant by an interfering factor is simply 'a factor that makes the law turn out to be false', the hedged claim says no more than 'the relation $s = ½ gt^2$ holds, unless it does not'. This is the second horn of the dilemma ("trivialty"). In what follows, we will call this 'Lange's dilemma' (named after Lange 1993, p. 235). The dilemma poses a challenge for an account of truth-conditions of cp-law statements.

4 Statistical Accounts of Special Science Laws

Earman and Roberts are quite pessimistic with regard to spelling out the truth conditions of cp-laws. However, this is not a major problem, they argue, because fundamental laws are not in need of cp-clauses and special science generalizations should not be understood as cp-laws either. Rather cp-laws play the scientific role of gesturing towards underlying generalizations that are more precise and not in need of cp-clauses:

> [A] ‘ceteris paribus law’ is an element of a ‘work in progress’, an embryonic theory on its way to being developed to the point where it makes definite claims about the world. It has been found that in a vaguely defined set of circumstances, a given generalizations has appeared to be *mostly right* or *mostly reliable*, and there is a hunch that somewhere in the neighborhood is a genuine, well-defined generalization, for which the search is on. (Earman and Roberts 1999, p. 466; emphasis added)

The essential point in this quote is that the *preliminary* formulation of a cp-law – that is “mostly right or mostly reliable” – belongs to the “context of discovery” of a search for a well-defined generalization. In the case of the special sciences, the result of the successful search for a well-defined generalization is a *statistical* generalization. By way of illustration Earman and Roberts refer to a case Kincaid discusses as an example of a statistical generalization: Jeffery Paiges’ study of revolutions in agrarian societies. Earman and Roberts discuss one of Paiges’ empirical findings as an example of a special science generalization: commercial hacienda systems tend to lead to agrarian revolt, whereas plantation systems tend to lead to labor reform (also mentioned in Roberts 2004, p. 165). Paiges argues for these claims on the basis of classifications (e.g. hacienda systems as opposed to other agrarian systems) and statistical analyses.

The statistical account permits two readings. According to the first and more liberal reading, Earman and Roberts reconstruct Paiges’ statistical generalization as follows: ‘*It is mostly true* that commercial hacienda systems lead to agrarian revolt, whereas plantation systems lead to labor reform.’ This mostly-statement is true, if it is the case that a generalizations holds in the majority of intended applications, i.e. if it is the case that in the majority of agrarian systems the generalization ‘if it is a commercial hacienda, then …’ holds. It is essential to this reading that a special science generalization is qualified by the operator ‘it is mostly true. For this reason we will call this reading of the statistical account the ‘mostly-reading’. It is worth stressing two points regarding this reading: firstly, a sentence of the form ‘it is mostly the case that p’ allows ‘p’ to be a *deterministic* as well as *statistical* a generalization. Secondly, Earman and Roberts claim that there is no need of a cp-clause. The clause has been replaced by ‘it is mostly right’. The non-strict character of the generalization is derived from the fact that the generalization does not hold in all (but the majority of) intended applications.

Elsewhere Earman and Roberts present their account of special science generalizations in slightly different words. Typical special science generalizations, they argue, are claims about “actual correlation among variables across various populations” (Earman and Roberts 1999, p. 467). These statements assert “a certain precisely defined statistical relation among well-defined variables” (Earman and Roberts 1999, p. 467, also Roberts 2004). That is, special science laws are *statistical generalizations* of the following form:

> In population H, a variable P is positively statistically correlated with variable S across all sub-populations that are homogeneous with respect to the variables $V_1, \ldots, V_n$ (Earman and Roberts 1999, p. 467).

This suggests that the above special science generalization should be reconstructed as follows: in all intended applications (i.e. all agrarian systems that are homogenous w.r.t. the values of the variables $V_1, \ldots, V_n$), there is a positive non-strict correlation between commercial hacienda systems and agrarian revolt, as well as between plantation systems and labor reform. This reading captures the non-strict character of special science generalizations by understanding these generalizations as statements about non-trivial conditional probabilities (i.e. conditional probabilities other than 0 and 1).[2] Let us call this reading of the statistical account the 'positive correlation-reading'. Again, Earman and Roberts claim that this reading of special science generalizations appears to dispense with cp-clauses.

In sum, the essential difference between the two readings is that the mostly-reading of special science generalizations is compatible with these generalizations being deterministic and probabilistic, while the positive-correlation-reading requires understanding special science generalizations as non-strict statistical generalizations. Both readings are intended to capture the 'non-strict' character of special science laws without making use of a lazy cp-clause.

In the recent literature, at least one other version of the statistical account has been advocated by Gerhard Schurz.[3] Schurz (2001, 2002) argues that special science laws ought to be understood as *normic* laws of the form 'normally, As are Bs'. What matters most for our present purposes is that normic laws imply what Schurz calls the statistical consequence thesis. The latter thesis consists in the assertion of "a high statistical probability of Ax conditional on Bx" (Schurz 2002, p. 365) or "numerically unspecified statistical generalizations of the form 'Most *A*s are *B*s'" (Reutlinger et al. 2011, Sect. 8.1).[4] Schurz's normic laws can be understood as an instance of the actual-correlation reading. Schurz's as well as Earman and Roberts' views have in common that they reconstruct special science generalizations, which appear to be hedged by a lazy cp-clauses, as statistical generalizations.

The statistical account of special science generalizations (both according to the mostly-reading and the positive-correlation reading) appears to be promising in at least three important respects:

1. Prima facie, a non-lazy cp-clause is not needed.
2. Statistical generalizations are indeed (dis)confirmable by evidence.
3. Statistical generalizations stating correlations capture the non-strict, non-universal, and exception-ridden character of generalizations in the special sciences.

[2] Probabilities are interpreted as actual frequencies here, for a discussion of this point see Reutlinger (manuscript).

[3] When characterizing dispositional terms, Rudolf Carnap already refers to an "escape clause" of the form "unless there are disturbing factors or provided the environment is in a normal state" and "usual circumstances in a laboratory" (Carnap 1956, p. 59).

[4] Schurz (2001, 2002) provides an evolution-theoretic argument for the statistical consequence thesis. A discussion of this argument would exceed the length of this paper (cf. also Reutlinger et al. 2011, Sect. 8.1). Instead we focus only on the conclusion (i.e. the normic account as a special case of the statistical account).

If the statistical account could be defended for special science generalizations the pay-off would indeed be considerable. We will, however, argue that this account does not work. It may be adequate for *some* special science generalizations but not as a general account of special science generalizations. In what follows we present two arguments against the statistical account. The first argument is directed against the mostly-reading. The second argument is directed against both readings.

5 Objection I: Cartwright's Dilemma

In this section, we primarily address the mostly-reading.[5] That is, we are interested in the claim that special science generalizations that appear to need a cp-clause, ought to be reconstructed as asserting that the generalization in question holds *mostly*.

We will start with a problem that Nancy Cartwright posed. The point of presenting the problem is not that those generalizations that can in fact be reconstructed according to the mostly-reading fall under the problem. Rather, the problem highlights the fact that there are many special science generalizations that cannot be reconstructed according to this reading in the first place. In her *How the Laws of Physics Lie*, Cartwright presents an argument whose main target is the covering law model of scientific explanation. However, the force of the argument carries over to the statistical account. The gist of the argument can be stated as a dilemma for cp-laws:

> Ceteris paribus generalizations, read literally without the 'ceteris paribus' modifier, are false. [. . .]. On the other hand, with the modifier the ceteris paribus generalizations may be true, but they cover only those few cases where the conditions are right. (Cartwright 1983, p. 45).

The horns of this dilemma are *falsity* and *restricted applicability*. Newton's first law is an example (again from physics – we will turn to examples from the special sciences shortly) which can be used to illustrate the dilemma:

> Every body continues in its state of rest or of uniform motion in a straight line, unless it is compelled to change that state by forces impressed upon it. (Newton 1999, p. 416)

Without the qualification "unless it is compelled to change that state by forces impressed upon it" Newton's first law is false. On the other hand, if the law is qualified by a cp-clause, then it applies to very few cases (if any cases at all). Call this dilemma: Cartwright's dilemma. It is worth noting that Cartwright's dilemma differs from Lange's dilemma, as the horns of the latter are *falsity* and *triviality* (see

[5] The argument also affects the positive-correlation-reading if the positive correlations in questions are *high* correlation and correlations are interpreted as actual frequencies.

Sect. 3). Unlike in the case of Lange's dilemma, Cartwright's point is not that cp-laws might be trivially true but rather that it is difficult to see why we should care about them if they cover only rarely occurring situations.

The dilemma is not a dilemma for those special science generalizations that might be adequately reconstructed according to the mostly-reading (we have not argued that there aren't any). The important point of the dilemma is that the mostly-reading cannot be a *general* account of special science laws. The dilemma highlights the fact that there are many generalizations, which appear to need a cp-clause (whether special science or not), because the generalizations cover only rare cases and can thus not be reconstructed as applying to most cases.[6] The important point for the goal of our paper is thus one of the premises of Cartwright's argument: There are many cp-laws (both in physics and in the special sciences) covering only very special *rarely occurring* situations.

Examples of generalizations in the special sciences that cover only rare cases are not far to seek. Consider two cases from economics: as we have stated before, economic agents maximize their expected utility. Rational agents are assumed to have complete information, transitive preferences etc. These features of agents are usually taken to be idealized because no real world agent has complete information. The law of demand holds under the condition of perfect competition. Perfect competition involves, among other things, perfectly informed agents that are competing and zero transaction costs. Idealized antecedent conditions or idealized conditions of application (such as 'if the population size is infinite ...', 'if mating occurs randomly ...') are also frequent in the case of generalizations in population ecology and evolutionary biology (Godfrey-Smith 2009; French 2011; Rice 2012). In analogy with Newton's first law, these examples suggest that laws in general should not be read as asserting that the relevant conditions of application obtain frequently. Cartwright's dilemma also applies to the examples from the special sciences: on the one hand, if we understand, say, the law of demand as a claim about what mostly happens, then the law would most certainly turn out to be false. On the other hand, if one qualifies the law by an exclusive cp-clause, then it does not apply to most real world cases.

It is worth pointing out that the problem of 'falsity and restricted applicability' is not genuine to exclusive cp-conditions. The problem might very well arise in the case of comparative cp-laws such as 'if the supply of a commodity increases (decreases), the price decreases (increases)'. As we have seen this statement is a comparative cp-law because (among other things) it requires that certain factors remain constant (e.g. demand). Should we read this cp-law as asserting that (among other things) the constancy of demand is a condition that obtains frequently, i.e. in most markets? It is unlikely – this assumption would presumably turn the law into a straightforward falsehood.

[6] We are not going to discuss a solution that succeeds in avoiding Cartwright's dilemma in this paper. See Hüttemann (1998) and (2012) for an attempted solution of this problem.

The conclusion we want to draw is that it is inadequate to reconstruct laws of the special sciences as claims about what mostly happens. While it may be true in some cases that a (deterministic or probabilistic) law statement holds in most intended applications, this cannot be a necessary condition for their truth or for their respectability.

One additional remark: we have objected to replacing cp-clauses by phrases such as "it is mostly true". However, the core of our objection is not concerned with the vagueness of "mostly". More importantly, we worry that very often whether or not a generalization is accepted as a cp-law does not at all depend on the frequency with which the relevant conditions are actualized. It is not in general part of the content of a special science law or generalization to state how often (whether characterized vaguely or quantitatively precise) its antecedent conditions are fulfilled (for a similar observation see Hempel 1988, Sect. 5).

6 Objection II: Lange's Dilemma and Non-Lazy cp-Conditions

Our second objection applies to the positive-correlation-reading and – a fortiori – to the mostly-reading too. According to the positive-correlation-reading, the *statistical* character of a special science generalization accounts for the exceptions – that is, a generalization has exceptions in the sense that it is a claim about non-trivial conditional probabilities. The following might be an illustration: in all agrarian societies a certain non-strict correlation (e.g. between a certain kinds of farming and kinds of political activities) holds – provided a certain finite number of conditions[7] obtains (stated as the claim that the variables $V_1, \ldots, V_n$ takes particular values). According to Earman and Roberts, one does not need a cp-clause because a non-strict correlation naturally allows for exceptions. However, it is difficult to see how this move could provide a solution to the problem of interpreting special science generalizations. In the remainder of this section, we provide an objection to the positive-correlation reading. This is our second objection to the statistical account of special science generalizations.

Our objection consists in the worry that the statistical account does not get rid of non-lazy cp-conditions. If this worry is justified, then the statistical account does not live up to Earman and Roberts's original aspirations of providing an account of special science laws that does not rely on non-lazy cp-conditions (see end of Sect. 4). The question we want to press is whether Earman and Roberts are justified to claim that the conditions can be considered to be *lazy* cp-conditions. To be precise, conditions are stated in terms of the variables in $\{V_1, \ldots, V_n\}$ taking a certain (range of) value(s). As mentioned in Sect. 2, cp-conditions are lazy if the list

[7] Conditions such as the "proximity of progressive urban political parties" (Earman and Roberts 1999, p. 468).

of conditions is *either* finite ('lazy$_1$') *or* finite and entirely in the scope of a special science ('lazy$_2$'). It is a striking fact that Earman and Roberts do not present an argument for the claim for the claim that the list of variables $V_1, \ldots, V_n$ and, thus, the corresponding list of conditions is finite (which is tantamount to the claim that the cp-conditions are lazy cp-conditions). We think they need an argument.

How do non-lazy cp-conditions enter the statistical account? Recall that Earman and Roberts refer to statistical generalizations that include other variables than the antecedent P: "in population H, a variable *P* is positively statistically correlated with variable *S* across all sub-populations that are homogeneous *with respect to the variables* $V_1, \ldots, V_n$" (Earman and Roberts 1999, p. 467, emphasis added). One way to describe the complex antecedent of this generalization is to say that even probabilistic generalizations are qualified by a comparative cp-clause (Schurz 2002; Reutlinger et al. 2011, Sect. 3.1). The comparative reading of the cp-clause specifies that, for instance, P and S are correlated if other variables $V_1, \ldots, V_n$ take specific values. The comparative reading corresponds to the literal translation of ceteris paribus, i.e. other things being equal.

So, given the comparative reading of cp-conditions, how does a *non-lazy* cp-clause enter the statistical account of special science laws? It is very plausible that the V_i of an economic or ecological statistical generalizations include physical and biological conditions that are not in the conceptual and methodological scope of the discipline in question. Consider two examples. *First,* rational agents are thought of as maximizing their utilities if they are not drugged. The condition of not being drugged might be part of the *implicit* knowledge of economist, but this condition is outside of the scope of standard micro-economics.[8]

Secondly, consider another illustration by Marc Lange. The area law of island-biogeography states:

> the equilibrium number S of a species of a given taxonomic group on an island (as far as creatures are concerned) increases [polynomially] with the islands area [A]: S = c×Az. The (positive-valued) constants c and z are specific to the taxonomic group and island group. (Lange 2002, p. 416f.)

Suppose we interpret the area law as a statistical generalization, as the statistical account requires. As Lange observes, the truth of the area law – even on a statistical reading – partly depends on conditions that lie outside of the scope of interest of island bio-geographers.

> There are counterfactual suppositions under which the fundamental laws of physics would still have held, but under which the 'area law' is not preserved. For example, had Earth lacked a magnetic field, then cosmic rays would have bombarded all latitudes, which might well have prevented life from arising, in which case [equilibrium number of a species] S would have been zero irrespective of [the island's area] A. (Lange 2002, p. 417)

[8] Similarly, Lange (2002) speaks of "off stage" variables, and Strevens (2008) refers to opaque conditions of application.

Lange says that the truth of the area law depends, among other things, on the actual strength of the magnetic field of the Earth. The law statement would be false if the magnetic field would be different than it actually is. Lange continues:

> The area law is not prevented from qualifying as an island-bio-geographical law [. . .] by its failure to be preserved under [this] [. . .] counterfactual supposition [. . .]. The supposition concerning Earth's magnetic field falls outside of island biogeography's range of interest. It twiddles with a parameter that island biogeography takes no notice of, or at least does not take it as a variable. (Lange 2002, p. 418)

In other words, the condition that the magnetic field of the Earth has its actual strength is a relevant condition, i.e. whether it obtains makes a difference to the truth or falsity of the area law. However, this condition is not salient in context of island biogeography.

What precisely do these examples show? *First*, they show that we have a good reason to speak of a non-lazy$_2$ cp-conditions that are relevant for statistical generalizations: that is, these conditions are not in the conceptual and methodological scope of the discipline in question and can thus not be captured by a statistical generalization formulated in the terminology of the special science in question. Thus, understanding special science laws as probabilistic generalizations does at least not replace non-lazy$_2$ cp-conditions.[9] However, even if these conditions are non-lazy$_2$ they need not be non-lazy$_1$. That is, even if the conditions cannot be stated in the vocabulary of the special science, there might be a finite list, if one allows for further conceptual and methodological resources (of other sciences). However, Earman and Roberts provide no argument for the claim that a finite list of such conditions will be available. Nor is there an argument for the claim that a finite list of conditions that fall *inside* the scope of the relevant special science can be given.

Secondly, if statistical special science laws have *either* (i) non-lazy$_2$-conditions that are at the same time non-lazy$_1$ (i.e. no finite list of them can be provided), *or* (ii) there are additional non-lazy$_1$ cp-conditions that fall inside the scope of the special science in question (if such a case is conceivable), then, we argue, the statistical approach cannot avoid Lange's dilemma. Suppose there is a non-lazy condition C for a higher-level statistical law 'p(B|A & V_1, . . . , V_n)=x', as sketched in the two examples above. *On the one hand,* if C is not added to the antecedent of the statistical law, then the statistical law is false. This is the first horn (falsity) of Lange's dilemma. *On the other hand*, if C is an open-ended list (non-lazy$_1$), then 'p(B|A & V_1, . . . , V_n & C)=x' becomes a statement without any clear meaning. According to Earman and Roberts' own reasoning, a law statement including an open-ended list of conditions C is in danger of becoming a trivial truth such as 'most As are Bs, unless something interferes' or 'A and B are correlated in conditions V_1, . . . , V_n*, unless something interferes*'.

To sum up the result of our second objection, the statistical account does not succeed in solving a problem it was designed for: it fails to dispense with non-lazy cp-conditions.

[9] Ironically for Earman and Roberts, Hempel (1988, p. 152f) argues for this point against Carnap.

We will conclude this section by discussing a possible objection to our argument. One might object that the statistical account captures nicely that – and even explains why – there are exceptions to a nomic relation. According to the statistical account, a higher-level statistical law simply describes a frequency that is the result of lazy and non-lazy interferers. One kind of referring to these interfering factors is to speak of 'noise' coming from the environment of the system under description.

We respond to this objection that it is true that in some cases these frequencies obtain and they can be explained by (environmental or lower-level) interfering factors (cf. Strevens 2008, Chap. 10, for an elaborate account of explaining frequencies in this way). However, as we have argued in Sect. 5, in the case of many laws there is not a good reason to believe that there are many of the frequencies required by the statistical account. Even if we focus on cases in which the relevant frequencies exist and in which the frequencies are the result of lower-level interfering or enabling factors, we would like to insist that the description of these lower-level factors is at least non-lazy$_2$ exercise. However, this is just what Earman and Roberts seem to deny.

7 Conclusion

Earman and Roberts advocate the statistical account of special science laws. At first glance, their account has the advantage of capturing the non-strict character of special science generalizations without being committed to allegedly problematic cp-conditions. We have presented two objections to the statistical account. The first objection attempts to establish the view that – contrary to the mostly-reading of the statistical account – it is not correct to say that special science generalizations should be interpreted as statements about what happens in most intended applications of the law. According to our second objection, the statistical account does not get rid of non-lazy cp-conditions. Hence, we conclude that the statistical account does not qualify as a general account of special science laws.

We will conclude with a brief outlook. If our arguments are sound, then statistical account does not succeed in dispensing with cp-conditions. This result motivates the following question: what is the positive account of lazy cp-conditions? Insofar as the authors are concerned, Hüttemann argues for a dispositionalist account of exclusive cp-laws. According to the dispositionalist 'cp, all *A*s are *B*s' is true if all *A*s have the disposition to *B*. In his (2012) Hüttemann argues that the two main objections against such an account can be countered, provided there exist laws of composition that describe how different dispositions contribute to one phenomenon. On this basis, he argues, it is (1) possible to account for the fact that referring to a disposition, which is not completely manifest, might nevertheless contribute to an explanation of an actual phenomenon. Furthermore (2), the laws of composition help to explain how we might gain evidence for how a system would behave in the absence of disturbing factors even if actual disturbing factors are present. The contribution of the disturbing factors can be calculated and – on the basis of the laws of composition – it can be 'subtracted'. This shows that at least some exclusive cp-laws are empirically testable. Reutlinger (2011) advocates an updated version of a completer account. This approach accounts for the truth conditions of a cp-generalization by relying on two essential concepts: (a) the concept of minimal invariance captures a relevance relation between the variables explicitly figuring in the generalization, and (b) the notion of a quasi-Newtonian law is used to describe the influence of disturbing factors.

References

Carnap, R. (1956). The methodological character of theoretical concepts. In H. Feigl & M. Scriven (Eds.), *The foundations of science and the concepts of psychology and psychoanalysis* (Minnesota studies in the philosophy of science, Vol. I, pp. 38–76). Minneapolis: Minnesota University Press.

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

Earman, J., & Roberts, J. (1999). Ceteris Paribus, there is no problem of provisos. *Synthese, 118*, 439–478.

Earman, J., Roberts, J., & Smith, S. (2002). Ceteris Paribus lost. *Erkenntnis, 57*, 281–301.

French, S. (2011). Shifting to structures in physics and biology: A prophylactic for promiscuous realism. *Studies in History and Philosophy of Biological and Biomedical Sciences, 42*, 164–173.

Godfrey-Smith, P. (2009). Models and fictions in science. *Philosophical Studies, 143*, 101–116.

Hempel, C. (1988). Provisoes: A problem concerning the inferential function of scientific theories. *Erkenntnis, 28*, 147–164.

Hüttemann, A. (1998). Laws and dispositions. *Philosophy of Science, 65*, 121–135.

Hüttemann, A. (2012). Ceteris-paribus-Gesetze in der Physik. In M. Esfeld (Ed.), *Philosophie der Physik* (pp. 390–410). Berlin: Suhrkamp.

Lange, M. (1993). Natural laws and the problem of provisos. *Erkenntnis, 38*, 233–248.

Lange, M. (2002). Who's afraid of Ceteris Paribus laws? or: How I learned to stop worrying and love them. *Erkenntnis, 52*, 407–423.

Newton, I. (1999). *The Principia: Mathematical Principles of Natural Philosophy* (I. B. Cohen & A. Whitman, Trans.). Berkeley: University of California Press.

Reutlinger, A. (2011). A theory of non-universal laws. *International Studies in the Philosophy of Science, 25*, 97–117.

Reutlinger, A. (manuscript). CP-Laws versus statistical laws – What's the Difference?

Reutlinger, A., Hüttemann, A., & Schurz, G. (2011). Ceteris Paribus laws. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*. http://plato.stanford.edu/archives/spr2011/entries/ceteris-paribus/.

Rice, C. (2012). Optimality explanations: A plea for an alternative approach. *Biology and Philosophy, 27*, 685–703.

Roberts, J. (2004). There are no laws in the social sciences. In C. Hitchcock (Ed.), *Contemporary debates in the philosophy of science* (pp. 168–185). Oxford: Blackwell.

Schurz, G. (2001). What is *Normal*? an evolution theoretic foundation of normic laws and their relation to statistical normality. *Philosophy of Science, 28*, 476–497.

Schurz, G. (2002). Ceteris Paribus laws: Classification and deconstruction. *Erkenntnis, 52*, 351–372.

Strevens, M. (2008). *Depth. An account of scientific explanation*. Cambridge: Harvard University Press.

Part IV
Philosophy of the Physical Sciences: Philosophy of Quantum Mechanics

How to Use Quantum Theory Locally to Explain EPR-Bell Correlations

Richard Healey

Abstract I sketch a pragmatist interpretation of quantum theory and show how to use it to explain EPR-Bell correlations consistently with relativity. Quantum theory is not a locally causal theory, not because it *violates* Bell's local causality condition, but because that condition is simply inapplicable to it. Any agent can use quantum theory to show why EPR-Bell correlations are to be expected. For space-like separated measurements of vertical/horizontal polarization of each photon from a pair in Bell state Φ^+, an agent's explanation of why the distant measurement outcome matches his own appeals neither to a preferred frame nor to any direct connection or influence between these events. Here, as elsewhere, quantum theory helps one explain an initially puzzling phenomenon not by locating it in a causal net but by showing why its occurrence is just what one should have expected in the circumstances.

1 Introduction

A consensus seems to be emerging on two points:

i. There is a persuasive argument from the *intuitive causality principle* that

> The direct causes (and effects) of events are near by, and even the indirect causes (and effects) are no further away than permitted by the velocity of light (Bell 1990, 2004, p. 239)

to the probabilistic independence condition(s) key to the proof of Bell's theorem.

R. Healey (✉)
Philosophy Department, University of Arizona, 213 Social Sciences,
Tucson, AZ 85721-0027, USA
e-mail: rhealey@email.arizona.edu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_16,

ii. There is an "apparently essential conflict between any sharp formulation [of quantum theory] and relativity." (Bell 2004, p.172)

Accepting (i) and (ii) means denying that we can use current quantum theory to satisfactorily explain EPR-Bell correlations, and commits one to seek a theory violating Bell's intuitive principle quoted in (i).

In opposition to this consensus, I sketch a pragmatist interpretation of current quantum theory[1] and show how we can use it to explain EPR-Bell correlations consistently with relativity. Quantum theory is not a locally causal theory, not because it *violates* Bell's local causality condition based on his intuitive causality principle, but because that condition is simply inapplicable to quantum theory. Any agent can use quantum theory to show just why EPR-Bell correlations are to be expected, whether the relevant measurement events are time-like or space-like separated. In a case of space-like separated measurements of vertical/horizontal polarization of each photon from a pair in Bell state Φ^+, an agent's explanation of why the distant measurement outcome always matches his own appeals neither to a preferred frame nor to any direct connection or influence between these events. Here, as elsewhere, quantum theory helps one explain an initially puzzling phenomenon not by locating it in a causal net but by showing why its occurrence is just what one should have expected in the circumstances.

2 Bell's Route from Local Causality to Factorizability

Immediately after stating and illustrating the intuitive causality principle quoted in (i), Bell continues

> The above principle of local causality is not yet sufficiently sharp and clean for mathematics. Now it is precisely in cleaning up intuitive ideas for mathematics that one is likely to throw out the baby with the bathwater. So the next step should be viewed with the utmost suspicion.

Bell's next step is to state a condition of local causality motivated by this intuitive principle.

Local Causality Condition (Bell 2004, pp. 239–240):

> A theory will be said to be locally causal if the probabilities attached to values of local beables in a space-time region 1 are unaltered by specification of values of local beables in a space-like separated region 2, when what happens in the backward light cone of 1is already sufficiently specified, for example by a full specification of all local beables in a space-time region 3.

Following Norsen (2011), Seevinck and Uffink (2010) (*SU*) recently presented a persuasive reconstruction of Bell's argument that no locally causal theory can account for the patterns of statistical correlation expected on the basis of quantum

[1] A more complete outline of this interpretation may be found in (Healey 2012).

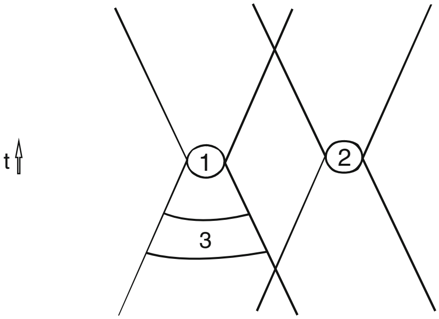


Fig. 1 Space-time diagram for Bell's *Local Causality*

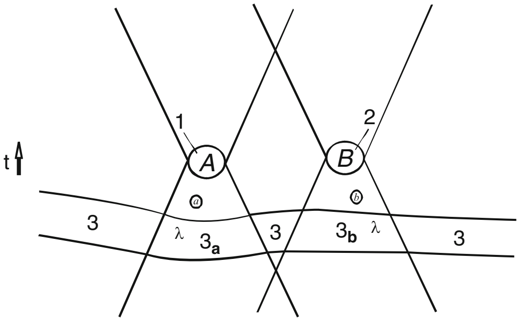


Fig. 2 Space-time diagram for Bell's proof

theory and now amply confirmed by experiments. They stress that Bell's condition of local causality is intended to apply to *theories* advanced as candidates for accounting for EPR-Bell correlations.

SU clarified Bell's notion of sufficiency as a combination of functional and statistical sufficiency, rendering the label b and random variable B (respectively) redundant for predicting $P_{a,b}(A|B,\lambda)$ – the probability a theory specifies for beable A representing the outcome recorded in region 1 given beables a, b representing the apparatus settings in regions 1,2 respectively, conditional on outcome B in region 2 and beable specification λ in a region 3 which smoothly joins 3_a to a similar slice 3_b right across the backward light cone of 2 so as to screen off both 1 and 2 from the overlap of their backward light cones. Figure 2 illustrates this situation.

In this situation, Bell's local causality condition implies

$$P_{a,b}\left(A\middle|B,\lambda\right) = P_a\left(A\middle|\lambda\right) \tag{1a}$$

By symmetry, interchanging '1' with '2', '*A*' with '*B*' and '*a*' with '*b*' implies

$$P_{a,b}\left(B\middle|A,\lambda\right) = P_b\left(B\middle|\lambda\right) \tag{1b}$$

SU offer Eqs. 1a and 1b as their mathematically sharp and clean formulation of the condition of local causality. Together, these equations imply the key factorizability condition

$$P_{a,b}\left(A,B\middle|\lambda\right)=P_a\left(A\middle|\lambda\right)\times P_b\left(B\middle|\lambda\right)\tag{2}$$

Factorizability used to derive CHSH inequalities. Experimental evidence that those inequalities are violated just as quantum theory leads one to expect is then taken to disconfirm the intuitive principle stated in point (i) and to dramatize point (ii).

SU endorse Bell's claim that orthodox quantum mechanics is not a locally causal theory, because it violates (Eq. 2). In the Bell state Φ^+, for example, the probability of recording a horizontally polarized photon in 1 depends on whether that polarization is recorded for the entangled photon in 2, since these records are perfectly correlated. The argument is that orthodox quantum theory specifies no beables in region 3 sufficient to render these outcomes probabilistically independent, as (Eq. 2) requires.

3 A Pragmatist Interpretation of Quantum Theory

Quantum theory is often taken to include the following interpretative principle:

Eigenstate-Eigenvalue Implication (EVI)	If a system's quantum state is an eigenstate of an observable, then that observable has the associated eigenvalue.

EVI implies that quantum theory specifies *some* beables in region 3 for a system in the Bell state Φ^+: the photon pair has a property associated with linear polarization, with respect to each axis in the relevant plane (*either* both along *or* both orthogonal, though *determinately* neither). Equation (2) fails even if one takes λ to be specified by these properties.

EVI may seem essential to quantum theory because without it the theory lacks descriptive content. But in an alternative view favored by many physicists, the quantum state's role is not descriptive, but merely predictive – to provide input to the algorithm specified by the Born Rule for calculating quantum probabilities. Bell and others have objected that this view reduces quantum theory to a blunt instrument for predicting statistics of measurement results obtained in ill-defined circumstances that is consequently unable to explain anything that happens outside the laboratory. Elsewhere I outline a pragmatist response to this objection and use it to indicate how quantum theory contributes to the explanation of natural phenomena, outside as well as inside the laboratory, thereby acquitting quantum theory so interpreted of the charge of instrumentalism.

This pragmatist interpretation rejects *EVI* and assigns the quantum state two roles. It plays its core role in the algorithm provided by the Born Rule for assigning quantum probabilities to magnitude claims of the form $A \in \mathbf{\Delta}$ about a system σ: *The value of A on σ lies in Δ*, where A is an observable, σ is a physical system and Δ is a Borel set of real numbers. But the significance of a claim $A \in \mathbf{\Delta}$ varies with the circumstances to which it relates. Accordingly, the quantum state plays a preliminary role by modulating the *content* of a magnitude claim by modifying its inferential relations to other claims. Only when a magnitude claim has sufficiently articulated content is an agent entitled to base credence on its Born probability.

Any application of quantum theory involves claims describing a physical situation. While it is considered appropriate to make claims about where individual particles are detected contributing to the interference pattern in a contemporary two-slit interference experiment, claims about through which slit each particle went are typically alleged to be "meaningless". In its preliminary role the quantum state offers guidance on the inferential powers, and hence the content, of descriptive claims of the form $A \in \mathbf{\Delta}$. The key idea here is that even assuming unitary evolution of the joint quantum state of system and environment, delocalization of system state coherence into the environment will typically render significant descriptive claims about experimental results and the condition of apparatus and other macroscopic objects by endowing these claims with enough content to license an agent to adopt epistemic attitudes toward them.

Quantum states are relational on this interpretation. The function of Born probabilities is to offer an agent authoritative advice on how to apportion degrees of belief concerning significant claims of the form $A \in \mathbf{\Delta}$ about a physical system which the agent is not currently in a position to check. It follows that a system does not have a unique quantum state. For when agents (actually or merely hypothetically) occupy relevantly different physical situations they should assign different quantum states to one and the same system, even though both quantum states and the Born probabilities to which they give rise are perfectly objective.

4 How This Helps Explain EPR-Bell Correlations

When we explain some natural phenomenon, we describe what happens to physical systems. According to the pragmatist interpretation sketched in Sect. 3, neither a quantum state nor the Born rule describes a physical system. In that sense, quantum theory by itself explains nothing. But quantum theory does grant an agent a wide license to entertain certain magnitude claims about physical systems, including some that agent is warranted in making on the basis of its "experience" (viz. records of experiments and/or observations). These include, but are by no means exhausted by, reports of measurement outcomes.

To use quantum theory to explain a regularity involving systems of certain types an agent begins by re-presenting that regularity in terms of warranted magnitude claims about these target systems. The next step is to use available information about

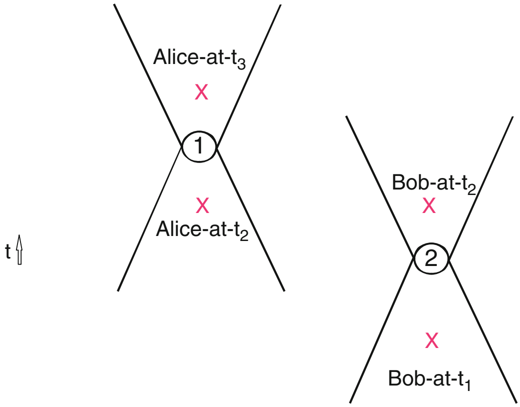


Fig. 3 Alice's and Bob's space-like separated polarization measurements

other physical systems in the spatiotemporal surroundings (based on magnitude claims about these systems) to assign a quantum state to the target system(s). These claims about neighboring systems describe the conditions on which the regularity depends: but they may concern what happens outside the causal past of some or all target systems.[2] The quantum state(s) assigned on the basis of these claims are objective even though they do not themselves either describe or represent properties or relations of the target systems. In this way a quantum state provides an objective mediation between physical conditions described in the (often implicit) *explanans* claims and the *explanandum* regularity which depends on them. But it does not describe a causal process that brings about instances of this regularity.

Quantum theory helps the whole scientific community explain a physical regularity by

(a) granting a wide license to warranted magnitude claims concerning events in the history of physical systems involved, either outside or inside the laboratory, and
(b) showing why the truth of some such claims was to be expected, given others on which they depend (by appeal to the Born rule).

But quantum theory may help differently situated agents explain the same phenomena differently.

Figure 3 is a space-time diagram from the perspective of Bob's inertial frame depicting space-like separated measurements by Alice and Bob in regions 1, 2 respectively on a photon pair in the Bell state Φ^+.

At t_1 each takes the polarization state of the L-R photon pair to be $|\Phi^+> = 1/\sqrt{2}\ (|HH> + |VV>)$. What warrants this quantum state assignment is their knowledge of the conditions under which the photon pair was produced –

[2] The delayed-choice entanglement-swapping experiment described in Sect. 4.5 of (Healey 2012) provides one illuminating example of this.

perhaps by parametric down-conversion of laser light by passage through a nonlinear crystal. Such knowledge depends on observation of the physical systems involved in producing the pair, and is expressible in claims about what Bell called "beables", including "the settings of switches and knobs on experimental equipment, the currents in coils, and the readings of instruments" (Bell 2004, p. 52). Then Alice measures polarization of photon L along axis a, Bob measures polarization of photon R along axis b. Decoherence at the photon detectors licenses both of them to base their credences, for each of the possible outcomes of Alice's polarization measurement, on probabilities given by the Born Rule.

At t_2, after recording polarization B for R, Bob ascribes state $|B>$ to L, and uses the Born Rule to calculate $P(A) = |<A|B>|^2$ for Alice to record polarization of L along the a-axis. At t_2, Alice ascribes state $\boldsymbol{\rho} = ½\mathbf{I}$ to L, and uses the Born Rule to calculate $P(A) = ½$ that she will record polarization of L along the a-axis. Each wisely uses the calculated probability to guide his or her expectations as to the outcome of Alice's measurement.

Alice's statistics of her outcomes in many repetitions of the experiment are just what her quantum state ½**I** for L led her to expect, thereby explaining her results. Bob's statistics for Alice's outcomes (in many repetitions in which his outcome is B) are just what his quantum state $| B>$ for L led him to expect, thereby explaining Alice's results.

There is no question as to which, if either, of the quantum states $| B>$, ½**I** was the *real* state of Alice's photon at t_2. The question as to which of the different probabilities $|<A|B>|^2$ or ½ gives the *real* "chance" at t_2 of Alice's outcome simply doesn't arise—even though neither Bob's nor Alice's Born probability is subjective. This discussion applies independent of the time-order in Alice's frame of the events 1, 2.

Probability in quantum theory is objective but relational. Born probabilities aren't local beables representing localized chances: they offer authoritative (different) advice to differently situated agents on what to expect, and thereby explain the statistical patterns each records. Quantum theory does not prescribe a single probability for Alice's (Bob's) measurement outcome in 1(2), but probabilities tailored to the local situation of each of them. This has implications for Bell's argument for *Factorizability* that began with his intuitive causality principle.

5 Causality and Relativity

The relational nature of quantum states and Born probabilities means that Bell's *Local Causality* condition cannot be applied to quantum theory, since that theory does not attach a unique probability (at a time) to each value of a relevant local beable in a later space-time region. Recall the mathematics Bell based on his "sharp and clean" condition for a theory to be locally causal:

$$P_{a,b}\left(A\middle|B,\lambda\right) = P_a\left(A\middle|\lambda\right) \tag{1a}$$

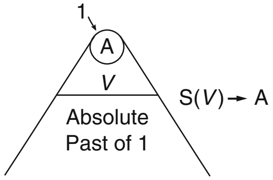


Fig. 4 Local determinism

Bell and others understand (Eq. 1a) to contain two expressions, each intended to represent a single local magnitude—*the* probability at t_2 of A in region 1. So understood, Eq. 1a simply can't be applied in quantum theory, since the theory denies there is any such magnitude. Alice's Born probability $P_a(A|\lambda)$ $(=P_{a,b}(A|\lambda))$ at t_2 cannot be identified with Bob's Born probability $P_{a,b}(A|\lambda)$ at t_2, since the different locations of Alice and Bob mean they should assign different quantum states to L at t_2. But if Eq. 1a is understood to equate two distinct magnitudes—Bob's best credence at t_2 and Alice's best credence at $t_2(P_a(A|\lambda))$—then no-one accepting quantum theory should expect these to be equal.[3]

If we follow Bell's own advice and view the step from his intuitive causality principle to his *Local Causality* condition with the suspicion it deserves, we see that by taking that step we implicitly exclude quantum theory from the class of theories advanced as candidates for accounting for EPR-Bell correlations. And so quantum theory remains immune from conclusions derived in the rest of the argument from *Local Causality* to *Factorizability*, and thence to Bell/CHSH inequalities.

This leaves three questions. Why did Bell consider his *Local Causality* condition an adequate explication of his intuitive causality principle, as applied to candidate theories? Is the quantum theoretic explanation of EPR-Bell correlations sketched in Sect. 4 really compatible with that intuitive causality principle? Can that explanation be made relativistically invariant?

Bell introduced an earlier version of his *Local Causality* condition in (Bell 1975) as a generalization of what he there called local determinism:

> In Maxwell's theory, the fields in any space-time region 1 are determined by those in any space region V, at some time t, which fully closes the backward light cone of 1 [Fig. 4]. Because the region V is limited, localized, we will say the theory exhibits *local determinism*. We would like to form some notion of *local causality* in theories which are not deterministic, in which the correlations prescribed by the theory, for the beables, are weaker. (Bell 2004, p. 53)

Bell thought of *local causality* as the appropriate generalization from a deterministic to a stochastic theory, as illustrated in Fig. 5.

[3] At t_2, Bob cannot apply the Born Rule to Φ^+ since R has been absorbed in his detector. But he can find a different use for the expression '$P_{a,b}(A|B,\lambda)$'. The conditional probability rule gives $P_{a,b}(A|B,\lambda) = P_{a,b}(A\,\&\,B|\lambda)/P_{a,b}(B|\lambda)$. Suppose we interpret these probabilities as Bob's best credences at t_2. At t_2, Bob is sure that his outcome is B, that his setting is b, and that Alice's

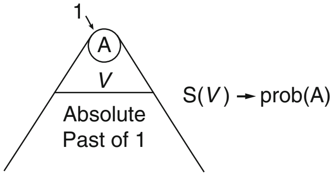


Fig. 5 Bell's stochastic generalization

This generalization treats the *probability* a theory assigns to a local beable as *itself* a local beable, uniquely correlated to a state in the (absolute) past of that assignment. It is not an appropriate generalization for a theory that assigns a probability to a local beable that is *relational* (though objective) and is therefore a function of the space-time location of an actual or hypothetical agent for whom it prescribes a rational degree of belief.

To check the compatibility with Bell's intuitive causality principle of the explanations offered in Sect. 4 using quantum theory, we need to determine their causal content. We saw that the quantum states appealed to in these explanations do not describe a causal process that brings about instances of the regularity explained. But whether or not there is such a process, in each instance of a Bell-EPR correlation there are three key events that might seem to be directly or indirectly causally related: the result of each polarization measurement, and the emission of the photon pair in the Bell state Φ^+. The case for a violation of Bell's intuitive causality principle is strongest when the polarization measurements produce macroscopically observable records in events in space-like separated regions 1, 2 of Fig. 3.

Even though the explanations offered in Sect. 4 include no explicit causal talk, they do underwrite counterfactual connections between Alice's and Bob's results: these are strict in case $a = b$, when quantum theory warrants Bob in claiming

If I had obtained a different polarization result, then so would Alice.

Counterfactual analyses of causation in the tradition of (Lewis 1973) remain popular among some philosophers. Butterfield (1992) used Lewis's own analysis to show that it implies that Bob's result caused Alice's in this situation. But quantum theory, interpreted as sketched in Sect. 3, does not warrant the claim that intervention at space-like separation from 1 could manipulate Alice's result.[4] Bell's causality principle is meant to be intuitive: there is nothing intuitive about a causal connection that would be useless for making things happen.

The quantum theoretic explanation of EPR-Bell correlations sketched in Sect. 4 was not relativistically invariant in so far as it implicitly assumed that the quantum state assignments were made in Bob's inertial frame. Compatibility with special

is a. So $P_{a,b}(B|\lambda) = 1$, $P_{a,b}(A \,\&\, B|\lambda) = P_{a,b}(A|\lambda)$. Hence $P_{a,b}(A|B,\lambda) = P_{a,b}(A|\lambda)$ is Bob's best credence at t_2.

[4] It is also a consequence of Woodward's (2003) more sophisticated interventionist account of causation that the counterfactual connections between Bob's and Alice's results should not be understood causally. Space does not permit me to demonstrate this here.

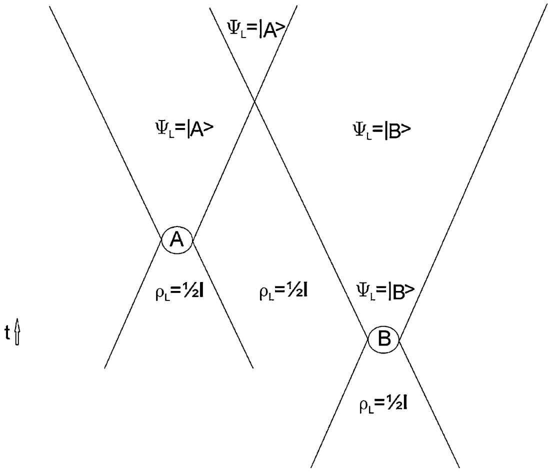


Fig. 6 L's polarization state at various space-time locations

relativity may be secured by requiring all quantum state assignments to be Lorentz covariant. If Alice is at rest in a different inertial frame, she may assign the appropriate Lorentz transform of the initial Bell state and subsequent quantum states. Figure 6 depicts the relevant states of Alice's photon in various space-time regions in one inertial frame. To ensure Lorentz covariance, each of these states should be taken to transform in the usual way under Lorentz transformations.

Correctly interpreted, predictions of quantum theory that violate Bell inequalities derived from (Eq. 2) provide no support for points (i) or (ii). Quantum theory's failure to conform to (Eq. 1a, 1b) and hence (Eq. 2) is no indication of a superluminal causal connection.

Acknowledgments This publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation. I thank Prof. A. Zeilinger and Prof. M. Aspelmeyer, my hosts at the Institute for Quantum Optics and Quantum Information of the Austrian Academy of Sciences under the Templeton Research Fellows Program "Philosophers-Physicists Cooperation Project on the Nature of Quantum Reality", and the other scientists in the Institute for many enlightening discussions. I acknowledge the support of the Perimeter Institute for Theoretical Physics under their sabbatical program during my visit in Fall 2009: research at the Perimeter Institute is supported by the Government of Canada through Industry Canada and by the Province of Ontario through the Ministry of Research and Innovation.

References

Bell, J. S. (1975). *The theory of local beables*. TH-2053-CERN, 28 July 2004, in Bell, pp 52–62.

Bell, J. S. (1990). La nouvelle cuisine. In A. Sarlemijn & P. Kroes (Eds.), *Between science and technology* (pp. 97–115). Amsterdam: North-Holland. In Bell (2004), 232–48.

Bell, J. S. (2004). *Speakable and unspeakable in quantum mechanics (revised edition)*. Cambridge: Cambridge University Press.

Butterfield, J. (1992). David Lewis meets John Bell. *Philosophy of Science, 59*, 26–43.

Healey, R. (2012). Quantum theory: A pragmatist approach. *British Journal for the Philosophy of Science, 63*, 729–771.

Lewis, D. K. (1973). Causation. *Journal of Philosophy, 70*, 556–567.

Norsen, T. (2011). *J.S. Bell's concept of local causality*. http://arxiv.org/PS_cache/arxiv/pdf/0707/0707.0401v3.pdf

Seevinck, M. P., & Uffink, J. (2011). Not throwing out the baby with the bathwater: Bell's condition of local causality mathematically 'sharp and clean'. In D. Dieks et al. (Eds.), *Explanation, prediction and confirmation* (pp. 425–450). Berlin: Springer.

Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.

Modal Interpretations and Consecutive Measurements

Juan Sebastián Ardenghi, Olimpia Lombardi, and Martín Narvaja

Abstract The correlations between the outcomes of consecutive measurements are one of those issues so deeply entrenched in the quantum knowledge of physicists that, in many cases, they use them to test the acceptability of any proposal of interpretation of the theory. The aim of the present article is to show the serious obstacles that modal interpretations face when trying to adequately account for those correlations, and to argue that the difficulties can be overcome without giving up the main modal theses if partial traces are dropped but the measuring apparatuses are taken into account.

1 Introduction

The practice of physics does not always follow the same way as the philosophical reflection on the meaning and the implications of the theories underlying that practice. This situation is particularly clear in the case of quantum mechanics. Most physicists are trained in the context of an "orthodox" view, which involves elements not always completely coherent with each other; nevertheless, in their everyday practice they make efficient use of quantum mechanics and perfectly know what results are expected to be obtained in particular experimental situations. For this reason, when an interpretation is offered to a professional physicist, he immediately contrasts it with certain paradigmatic examples belonging to his strongly rooted quantum knowledge. The correlations between the outcomes of consecutive measurements are one of those examples. The aim of the present

J.S. Ardenghi • O. Lombardi (✉) • M. Narvaja
CONICET – Universidad de Buenos Aires, Crisólogo Larralde 3440, 6D (1430)
Ciudad de Buenos Aires, Argentina
e-mail: jsardenghi@gmail.com; olimpiafilo@arnet.com.ar; martinnarvaja@hotmail.com

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_17,

article is to show the challenge that consecutive measurements pose to modal interpretations, and to explain how this challenge can be overcome without giving up the main modal theses.

2 Consecutive Measurements

Let us consider a first kind measurement under the standard von Neumann model, according to which a quantum measurement is an interaction between a system S and a measuring apparatus M. Before the interaction, M is prepared in a ready-to-measure state $|r_0\rangle$, eigenvector of the pointer observable R of M, and the state of S is a superposition of the eigenstates $|a_i\rangle$ of an observable A of S. The interaction introduces a correlation between the eigenstates $|a_i\rangle$ of A and the eigenstates $|r_i\rangle$ of R:

$$|\psi_0\rangle = \sum_i c_i \, |a_i\rangle \otimes |r_0\rangle \quad \rightarrow \quad |\psi\rangle = \sum_i c_i \, |a_i\rangle \otimes |r_i\rangle \tag{1}$$

The measurement problem consists in explaining why, the state $|\psi\rangle$ being a superposition of the $|a_i\rangle \otimes |r_i\rangle$, the pointer R acquires a definite value.

A further question is what result is obtained when a second measurement of the same observable is performed on the same system. As it is well known in practice:

> When we measure a real dynamical variable ξ, [...] if we make a second measurement on the same dynamical variable ξ immediately after the first, the result of the second measurement must be the same as that of the first. Thus, after the first measurement has been made, there is no indeterminacy in the result of the second. (Dirac 1958, p. 36)

A paradigmatic example is given by consecutive Stern-Gerlach experiments: no matter what interpretation of the theory one uses to explain them, any physicist knows that, if the particle outcoming the first magnet is deflected in the $+z$-direction ($-z$-direction), a second magnet in the same direction will deflect it again in the $+z$-direction ($-z$-direction) with certainty. The non-trivial correlations between the outcomes of consecutive measurements of different observables are also well known: if the particle is deflected by the first magnet in the $+z$-direction ($-z$-direction), a second magnet in the x-direction will deflect it in the $+x$-direction or in the $-x$-direction with a probability of ½ for each case. However, although there is no doubt about the phenomenon, different explanations still coexist.

3 Collapse and Ensemble Interpretations

According to the collapse interpretation, in each particular detection, the pure state $|\psi\rangle$ of Eq. 1 indeterministically "collapses" to a component of the superposition, say, $|a_k\rangle \otimes |r_k\rangle$, which is interpreted as saying that the system S is in the state $|a_k\rangle$

and the apparatus M is in the state $|r_k\rangle$ with certainty; then, the observable A of S and the pointer R of M acquire the definite values a_k and r_k with certainty, respectively.

The collapse hypothesis gives a straightforward account of the agreement between the outcomes of consecutive measurements of the same observable. In fact, a second measurement of the observable A on the system S will maintain the system in the same state $|a_k\rangle$ resulting from the collapse in the first measurement. This means that the conditional probability of obtaining the value $r_k{}^{(2)}$ of the pointer $R^{(2)}$ of the apparatus $M^{(2)}$ in the second measurement, given that the value $r_k{}^{(1)}$ of $R^{(1)}$ of $M^{(1)}$ was obtained in the first measurement, is $pr(r_k{}^{(2)}/r_k{}^{(1)}) = 1$.

The collapse hypothesis also supplies a simple explanation to the non-trivial correlations between the outcomes of consecutive measurements of different observables (see, e.g., Cohen-Tannoudji et al. 1977). Let us consider a Stern-Gerlach experiment, where the observable A to be measured is the spin S^z in z-direction, with eigenvectors $|z_+\rangle$ and $|z_-\rangle$, and the role of the pointer R is played by the particle's momentum P^z in z-direction, with eigenvectors $|p_0{}^z\rangle$, $|p_+{}^z\rangle$ and $|p_-{}^z\rangle$. In general, the first interaction leads to a superposition $|\psi\rangle = c_+|z_+\rangle \otimes |p_+{}^z\rangle + \; c_-|z_-\rangle \otimes |p_-{}^z\rangle$; so, the first measurement collapses it to, say, $|z_+\rangle \otimes |p_+{}^z\rangle$ with certainty. In this case, the outcome of a second measurement of S^z will be $|p_+{}^z\rangle$ with certainty. But if the second measurement is performed on the spin S^x in x-direction, the state $|z_+\rangle$ must be expressed in the eigenbasis $\{|x_+\rangle,|x_-\rangle\}$ of the new observable,

$$|z_+\rangle = {}^1\!\big/\!_{\sqrt{2}} \; (|x_+\rangle + |x_-\rangle) \tag{2}$$

and the interaction correlates $|x_+\rangle, |x_-\rangle$ with the eigenvectors $|p_+{}^x\rangle, |p_-{}^x\rangle$ of the momentum P^x in x-direction, respectively. Then, the second measurement collapses the second superposition to, say, $|x_+\rangle \otimes |p_+{}^x\rangle$ with probability ½. Therefore, the conditional probability of obtaining the value $p_+{}^x$ of the momentum P^x in x-direction in the second measurement, given that the value $p_+{}^z$ of the momentum P^z in z-direction was obtained in the first measurement, is ½.

Whereas the collapse interpretation assumes that a pure state provides a complete and exhaustive description of an individual system, according to the ensemble interpretation a pure state describes the statistical properties of an ensemble of similarly prepared systems. From this perspective, the fact that the pure state $|\psi\rangle = \sum_i c_i |a_i\rangle \otimes |r_i\rangle$ of Eq. 1 is a superposition is not an obstacle to explain the definite value of the pointer R, since $|\psi\rangle$ describes an ensemble where the different $|a_i\rangle \otimes |r_i\rangle$ are distributed with their corresponding frequencies $|c_i|^2$ in the ensemble. Therefore, in each particular measurement we will necessarily detect a particular value a_k of A and a particular value r_k of R: probabilities measure our ignorance about the precise result to be obtained.

The ensemble interpretation also offers a simple account of the agreement of the outcomes of consecutive measurements by means of the idea of filtering-type measurement. In a measurement of the filtering-type, the ensemble of the systems coming from the first interaction, and represented by a superposition of the $|a_i\rangle \otimes |r_i\rangle$ (see Eq. 1), is firstly separated into i subensembles according to the value

r_i of the pointer R. The subsequent filtering process has the effect of removing all the elements of the superposition with values of R different than a particular value, say, r_k (see Ballentine 1998). The second measurement, then, is performed on only the obtained a_k -subensemble, now represented by the state $|a_k\rangle$. Since all the members of the new ensemble have the value a_k of the observable A, a second measurement of the observable A will give the same value of the pointer as in the first measurement. In a certain sense, the ensemble interpretation modifies the original understanding of the collapse: from being a non-unitary, stochastic, entropy-increasing physical process induced by measurements (von Neumann 1932), collapse becomes an epistemic fact resulting from the decrease of our ignorance when the ensemble is adequately restricted by a filtering process.

The point to stress here is the common feature shared by the collapse and the ensemble interpretations: in both cases, the definite value of the pointer is explained in terms of a *change* in the state emerging from the interaction. As we will see, the lack of this feature is precisely the obstacle that modal interpretations must overcome in their account of consecutive measurements.

4 Modal Interpretations as Single-System No-Collapse Interpretations

Although collapse and ensemble interpretations supply reasonable answers to the measurement problem, they must face several difficulties with respect to other interpretative issues. In a certain sense, modal interpretations attempt to overcome those difficulties from a realist and modal perspective. The main idea behind them is that quantum states constrain possibilities rather than actualities: "*the state delimits what can and cannot occur, and how likely it is −it delimits possibility, impossibility, and probability of occurrence− but does not say what actually occurs*" (van Fraassen 1991, p. 279). In spite of the differences among them, all modal interpretations agree on two points relevant to our discussion (see Dieks and Vermaas 1998; Lombardi and Dieks 2013):

- A quantum measurement is an ordinary physical interaction: there is *no collapse*.
- The quantum state refers to a *single system*, not to an ensemble of systems.

On the basis of these shared features, one can immediately notice the difficulty that modal interpretations face when the task is to explain the agreement between the outcomes of consecutive measurements of the same observable. On the one hand, since the quantum state refers to a single system, that agreement cannot be explained as the result of the partition of an original ensemble into subensembles, as in the ensemble interpretation. On the other hand, since the quantum state always evolves unitarily, the explanation of the outcome agreement can neither be based on collapse, as in the collapse interpretation.

From a general perspective, the difficulty is rooted in the fact that, according to modal interpretations, the quantum state always evolves according to the Schrödinger equation: measurement does not cause a non-unitary modification of the state. But the non-unitary step cannot be simply omitted when the task is to account for the correlations between the outcomes of consecutive measurements. In fact, the professional physicist, trained in the framework of the collapse interpretation —or, sometimes, of the ensemble interpretation— immediately notices that an always-unitary evolution may lead to wrong conclusions.[1] Let us consider a Stern-Gerlach measurement of spin S^z, as presented in Sect. 3. No matter how the definite reading of z_+ in the first measurement is explained, if there were no non-unitary process —our physicist friend would argue—, the measured system would continue in the original superposition $|\psi_0\rangle$, which, when expressed in the basis of S^x, would read

$$|\psi_0\rangle = c_+\,|z_+\rangle + \; c_-\,|z_-\rangle = {}^1\!\Big/_{\sqrt{2}}\,(c_+ + \; c_-)\,|x_+\rangle + {}^1\!\Big/_{\sqrt{2}}\,(c_+ - \; c_-)\,|x_-\rangle \tag{3}$$

Therefore, given that z_+ was obtained in the first measurement, the conditional probabilities $pr(z_+/z_+)$ and $pr(x_+/z_+)$ would be

$$pr\,(z_+/z_+) = |c_+|^2 \quad \text{and} \quad pr\,(x_+/z_+) = \frac{|c_+ + \; c_-|^2}{2} \tag{4}$$

which are notably different than the correct values 1 and 1/2, respectively, explained by the collapse and the ensemble interpretations in terms of the non-unitary step introduced by measurement. Moreover, the results of Eq. 4 would have been the same if the value z_-, instead of z_+, had been detected in the first measurement:

$$pr\,(z_+/z_+) = pr\,(z_+/z_-) = pr\,(z_+) \quad \text{and} \quad pr\,(x_+/z_+) = pr\,(x_+/z_-) = pr\,(x_+) \tag{5}$$

This means that the outcomes of the first and the second measurements would turn out to be independent to each other, completely at odds with the expected result.

5 Modal Interpretations and Consecutive Measurements

In this section we will consider the accounts of quantum measurement given by two modal interpretations: the Kochen-Dieks interpretation (KDI) (Kochen 1985; Dieks 1988, 1989, 1994) —or its generalization to mixed states, given by the

[1] This was the reaction of Rodolfo Gambini when we explained him the basics of modal interpretations. We are grateful to him for the interesting discussion that followed.

Vermaas-Dieks interpretation (VDI) (Vermaas and Dieks 1995)−, and the modal-Hamiltonian interpretation (MHI) (Lombardi and Castagnino 2008; Ardenghi et al. 2009; Lombardi et al. 2010; Ardenghi and Lombardi 2011). We have chosen these two cases because, although both "modal", they differ in the selection of the preferred context, that is, the set of the actual and definite-valued observables of the system: whereas the KDI/VDI's preferred context depends on the instantaneous state and, then, changes over time as the state evolves, in the MHI the preferred context depends on the system's Hamiltonian and, as a consequence, is time-invariant. We will see that, in spite of this difference, both face the same difficulty in the account of consecutive measurements.

The KDI exploits the biorthogonal (Schmidt) decomposition theorem to define the preferred context. In fact, if the state of the composite system after the correlation is $|\psi\rangle = \sum_i c_i |a_i\rangle \otimes |r_i\rangle$ (see Eq. 1), then the preferred context of the measured system S is defined by the set $\{|a_i\rangle\}$ and the preferred context of the measuring apparatus M is defined by the set $\{|r_i\rangle\}$. In the case of the VDI, the preferred context is given by the spectral resolution of the reduced state of the system, obtained by partial trace. This means that, in the measurement situation, where the reduced states of S and M are

$$\rho^r_S = Tr_{(M)}\,|\psi\rangle\,\langle\psi| = \sum_i |c_i|^2\,|a_i\,\rangle\,\langle a_i| \quad \text{and} \quad \rho^r_M = Tr_{(S)}\,|\psi\rangle\,\langle\psi| = \sum_i |c_i|^2\,|r_i\,\rangle\,\langle r_i| \tag{6}$$

the preferred context of S is defined by the projectors $|a_i\rangle\langle a_i|$ and the preferred context of M is defined by the projectors $|r_i\rangle\langle r_i|$. Therefore, in both KDI and VDI, the observables A of S and R of M acquire actual definite values, whose probabilities are given by the diagonal elements $|c_i|^2$ of the diagonalized reduced states.

The MHI endows the Hamiltonian of the closed system with a central role: the preferred context of a system is constituted by its total Hamiltonian H and all the observables commuting with H and having, at least, the same symmetries −degeneracies− as H. Then, in the last stage of measurement, when S and M do not longer interact with each other, the observable that defines the preferred context of M is its Hamiltonian H_M. Therefore, the commutation condition $[H_M, R] = 0$, which makes possible the reading of the pointer R by guaranteeing its stability, is also what explains the actual definite value of the pointer, even in non-ideal measurements. The probabilities corresponding to the different values of A and of R are given by the diagonal elements of the corresponding diagonalized reduced states $\rho_S{}^r$ of S and $\rho_M{}^r$ of M, as in KDI and VDI (see Eq. 6).

In both cases, KDI/VDI and MHI, reduced states are conceived as the quantum (mixed) states of the measured system S and of the measuring apparatus M after the interaction. In the case of the KDI/VDI, this assumption has no restrictions: $\rho_S{}^r$ and $\rho_M{}^r$ are the quantum states of S and of M, respectively, no matter whether M is still interacting with S or even with its large environment or not. As Dieks asserts,

> The projection operator $|\psi\rangle\langle\psi|$ is an observable of the total system [...]. But we are really interested in the individual properties of device and object taken by themselves.[...] we took [the reduced operator] W_1 for the state of system 1. This is standard practice in quantum mechanics; however, the usual justification relies on the probabilistic interpretation of the theory and the Born rule (Dieks 2007, pp. 298–299).

The MHI is more restrictive in this regard: $\rho_S{}^r$ and $\rho_M{}^r$ can be considered the quantum states of S and of M just in the case that there is no further interaction between S and M and, as a consequence, both states evolve unitarily according to the Schrödinger equation in its von Neumann version. In fact, according to the MHI, a quantum system is composite only when the subsystems do not interact:

> In this case, the initial states $\rho_0{}^1$ [...] and $\rho_0{}^2$ [...] of [the subsystems] S^1 and S^2, respectively, are obtained as $\rho_0{}^1 = Tr_{(2)}\ \rho_0$ and $\rho_0{}^2 = Tr_{(1)}\ \rho_0$, where $Tr_{(i)}\ \rho_0$ is a partial trace of [the initial state of the whole composite system] ρ_0, that is, the operation that traces over the Hilbert space of S^i (Lombardi and Castagnino 2008, p. 389).

It is clear that, independently of the difference between KDI/VDI and MHI with respect to the selection of the preferred context, in both cases, the reduced operators are conceived as representing the quantum states of the subsystems of a composite system.

In spite of its apparent naturalness and the good results that it supplies in the case of a single measurement, the assumption that the reduced operator represents the quantum state of a system is not innocuous when consecutive measurements are considered (see a seminal discussion in van Fraassen 1972). Let us take seriously that, in the case of consecutive measurements, the reduced operators

$$\rho_S^{r(1)} = Tr_{\left(M^{(1)}\right)}\left|\psi^{(1)}\right\rangle\left\langle\psi^{(1)}\right| = \sum_i |c_i|^2 \left|a_i\right\rangle\left\langle a_i\right|,$$

$$\rho_M^{r(1)} = Tr_{(S)}\left|\psi^{(1)}\right\rangle\left\langle\psi^{(1)}\right| = \sum_i |c_i|^2 \left|r_i^{(1)}\right\rangle\left\langle r_i^{(1)}\right| \qquad (7)$$

really represent the quantum states of the measured system S and of the first measuring system $M^{(1)}$ respectively. In spite of the differences between them, according to both KDI/VDI and MHI the pointer $R^{(1)}$ of $M^{(1)}$ is a definite-valued observable which, as a consequence, indeterministically acquires a certain value, say $r_k{}^{(1)}$, with its corresponding probability. But now the reduced state $\rho_S{}^{r(1)}$ is the state of the system before the second measurement, when the second measuring apparatus $M^{(2)}$ is in the ready-to measure state $|r_0{}^{(2)}\rangle$. So, a new correlation is established when the system interacts with the second apparatus:

$$\rho_0^{(2)} = \rho_S^{r(1)} \otimes \left|r_0^{(2)}\right\rangle\left\langle r_0^{(2)}\right| \quad \rightarrow \quad \rho^{(2)} = \sum\nolimits_i |c_i|^2 \left|a_i\right\rangle\left\langle a_i\right| \otimes \left|r_i^{(2)}\right\rangle\left\langle r_i^{(2)}\right| \quad (8)$$

Then, when that second interaction ends, the analogous account of measurement applies: $\rho_S{}^{r(2)}$ is the reduced operator corresponding to the system and $\rho_M{}^{r(2)}$ is the reduced operator corresponding to the second measuring apparatus,

$$\rho_S^{r(2)} = Tr_{(M^{(2)})}\rho^{(2)} = \sum_i |c_i|^2 \left|a_i\right\rangle \left\langle a_i\right| \quad \text{and}$$
$$\rho_M^{r(2)} = Tr_{(S)}\rho^{(2)} = \sum_i |c_i|^2 \left|r_i^{(2)}\right\rangle \left\langle r_i^{(2)}\right| \tag{9}$$

Then, the pointer $R^{(2)}$ of the second measuring apparatus $M^{(2)}$ indeterministically acquires a certain value, say $r_l{}^{(2)}$, with its corresponding probability, but with no relation with the value $r_k{}^{(1)}$ acquired by the pointer $R^{(1)}$ of the first measuring apparatus $M^{(1)}$.

Summing up, if we took seriously the assumption that the reduced operators are the legitimate quantum states of the measuring apparatuses, we have no way of explaining the agreement between the readings of both pointers, since the partial traces have cancelled the correlations between the values acquired by these two observables.

6 Consecutive Measurements Without Collapse

It is not difficult to imagine that, for our physicist friend, the challenge that consecutive measurements poses to modal interpretations is a good enough reason to discard them. Furthermore, he might even argue that the results just obtained are the best proof of the fact that quantum measurements cannot be correctly explained without the collapse hypothesis. However this is a too hasty conclusion. We will show that the correlations between the outcomes of consecutive measurements can be easily accounted for without collapse when partial traces are dropped but the measuring apparatuses are taken into account.

Let us consider again the case of consecutive Stern-Gerlach measurements as described in Sect. 3. After the first measurement, the correlated state is $|\psi^{(1)}\rangle = c_+|z_+\rangle \otimes |p_+{}^z\rangle + c_-|z_-\rangle \otimes |p_-{}^z\rangle$, and the second S^x-measuring apparatus is in its ready-to measure state $|p_0{}^x\rangle$:

$$\left|\psi_0^{(2)}\right\rangle = \left|\psi^{(1)}\right\rangle \otimes |p_0^x\rangle = \left(c_+|z_+\rangle \otimes \left|p_+^z\right\rangle + c_-|z_-\rangle \otimes |p_-^z\rangle\right) \otimes |p_0^x\rangle \tag{10}$$

In order to show the correlations to be introduced by the second measurement, the states $|z_+\rangle$ and $|z_-\rangle$ must be expressed in the basis $\{|x_+\rangle, |x_-\rangle\}$ of S^x:

$$|z_+\rangle = {}^1\!\big/\!_{\sqrt{2}}\ (|x_+\rangle + |x_-\rangle) \qquad |z_-\rangle = {}^1\!\big/\!_{\sqrt{2}}\ (|x_+\rangle - |x_-\rangle) \tag{11}$$

Then,

$$\left|\psi_0^{(2)}\right\rangle = \Big[\left(1/\sqrt{2}\right)\left(c_+\left|p_+^z\right\rangle + c_-|p_-^z\rangle\right) \otimes |x_+\rangle$$
$$+ \left(1/\sqrt{2}\right)\left(c_+\left|p_+^z\right\rangle - c_-|p_-^z\rangle\right) \otimes |x_-\rangle\Big] \otimes |p_0^x\rangle \tag{12}$$

Since there is no collapse, the second interaction with a magnetic field in x-direction establishes the correlation between $|\psi_0{}^{(2)}\rangle$ and the eigenstates $|p_+{}^x\rangle$ and $|p_-{}^x\rangle$ of the second pointer P^x:

$$\left|\psi^{(2)}\right\rangle = \left(1/\sqrt{2}\right)\left(c_+\left|p_+^z\right\rangle + c_-\left|p_-^z\right\rangle\right) \otimes |x_+\rangle \otimes \left|p_+^x\right\rangle + \\ + \left(1/\sqrt{2}\right)\left(c_+\left|p_+^z\right\rangle - c_-\left|p_-^z\right\rangle\right) \otimes |x_-\rangle \otimes \left|p_-^x\right\rangle \tag{13}$$

or equivalently,

$$\left|\psi^{(2)}\right\rangle = \left(c_+/\sqrt{2}\right)\ |x_+\rangle \otimes \left|p_+^z\right\rangle \otimes \left|p_+^x\right\rangle + \left(c_-/\sqrt{2}\right)\ |x_+\rangle \otimes \left|p_-^z\right\rangle \otimes \left|p_+^x\right\rangle + \\ + \left(c_+/\sqrt{2}\right)\ |x_-\rangle \otimes \left|p_+^z\right\rangle \otimes \left|p_-^x\right\rangle - \left(c_-/\sqrt{2}\right)\ |x_-\rangle \otimes \left|p_-^z\right\rangle \otimes \left|p_-^x\right\rangle \tag{14}$$

Now we can compute any conditional probability by combining the outcomes of the first and the second measurement. For instance, we can compute the probability that the second pointer –here the momentum P^x in x-direction– acquires the value $p_+{}^x$, given that the first pointer –here the momentum P^z in z-direction– acquired the value $p_+{}^z$, as follows:

$$pr\left(p_+^x/p_+^z\right) = \frac{pr\left(p_+^x \wedge p_+^z\right)}{pr\left(p_+^z\right)} \tag{15}$$

The crucial point to stress here is that $pr(p_+{}^x \wedge p_+{}^z)$ is a legitimate probability since the pointers P^x and P^z are commuting observables (see Laura and Vanni 2008). In this particular case $[P^x, P^z] = 0$ because they are orthogonal components of the momentum of a particle. But, in general, the pointers $R^{(1)}$ and $R^{(2)}$ of consecutive measurements commute because they belong to different measuring apparatuses which, as a consequence, are represented by different Hilbert spaces. So, from the first term of Eq. 14 we obtain

$$pr\left(p_+^x \wedge p_+^z\right) = |c_+|^2/2 \tag{16}$$

In turn, $pr(p_+{}^z)$ can be computed from the first and the third terms of Eq. 14:

$$pr\left(p_+^z\right) = \left(|c_+|^2/2\right) + \left(|c_+|^2/2\right) = |c_+|^2 \tag{17}$$

Therefore,

$$pr\left(p_+^x/p_+^z\right) = \frac{pr\left(p_+^x \wedge p_+^z\right)}{pr\left(p_+^z\right)} = \frac{|c_+|^2/2}{|c_+|^2} = \frac{1}{2} \tag{18}$$

as expected.

Of course, we could also decide to measure in the second measurement the same observable as in the first measurement. In this case, the basis rotation introduced in Eq. 12 is not necessary, and the second interaction leads the system to the state

$$\left|\psi^{(2)}\right\rangle = c_+ \left|z_+\right\rangle \otimes \left|p^z_+\right\rangle \otimes \left|p^z_+\right\rangle + c_- \left|z_-\right\rangle \otimes \left|p^z_-\right\rangle \otimes \left|p^z_-\right\rangle \tag{19}$$

It is clear that now the conditional probabilities are

$$\begin{aligned} pr\left(p^z_+/p^z_+\right) &= \frac{pr\left(p^z_+ \wedge p^z_+\right)}{pr\left(p^z_+\right)} = \frac{|c_+|^2}{|c_+|^2} = 1 \quad \text{and} \\ pr\left(p^z_-/p^z_+\right) &= \frac{pr\left(p^z_- \wedge p^z_+\right)}{pr\left(p^z_+\right)} = \frac{0}{|c_+|^2} = 0 \end{aligned} \tag{20}$$

as expected. These results can be generalized for any different observables and for any number of consecutive measurements. As a consequence, and by contrast to what our physicist friend might believe, the experimental observed correlations between the outcomes of consecutive measurements can be perfectly explained with no need of the collapse hypothesis.

7 Conclusions and Perspectives

As we have seen, modal interpretations face serious obstacles to adequately account for correlations between the outcomes of consecutive measurements. However, when the explanation of consecutive measurements as presented in the previous section is compared with that of Sect. 5, the difference between them turns out to be clear: in the failed account a partial trace was introduced as the operation that defines the states of the measured system and the measuring apparatus after the first interaction. This means that the difficulties that modal interpretations face are not the result of the rejection of the collapse hypothesis, but a consequence of endowing reduced operators with a feature alien to them, namely, that of representing quantum states.

Our considerations do not imply the uselessness of the reduced operators in quantum mechanics. In fact, reduced operators are commonly used in practice to describe open systems, as in the case of the theory of decoherence (Zurek 2003). However, one must always recall the well-known distinction between proper and improper mixtures (d'Espagnat 1976): reduced states are improper mixtures resulting from partial trace and, as a consequence, they cancel correlations, which, although unobservable in many cases, are empirically manifested as measurable conditional probabilities in consecutive measurements. So, these results open up two lines of research for future work. One of them is the analysis of the status of reduced states in the context of quantum mechanics and the possibility of considering them as coarse-grained states. The other is the reconsideration of the role played by reduced operators in the interpretative context, a task that may lead to a better understanding of the holistic features of quantum mechanics.

References

Ardenghi, J. S., Castagnino, M., & Lombardi, O. (2009). Quantum mechanics: Modal interpretation and Galilean transformations. *Foundations of Physics, 39*, 1023–1045.

Ardenghi, J. S., & Lombardi, O. (2011). The modal-Hamiltonian interpretation of quantum mechanics as a kind of "atomic" interpretation. *Physics Research International, 2011*, 379604.

Ballentine, L. (1998). *Quantum mechanics: A modern development*. Singapore: World Scientific.

Cohen-Tannoudji, C., Diu, B., & Lalöe, F. (1977). *Quantum mechanics*. New York: Wiley.

d'Espagnat, B. (1976). *Conceptual foundations of quantum mechanics*. Reading: Benjamin.

Dieks, D. (1988). The formalism of quantum theory: an objective description of reality? *Annalen der Physik, 7*, 174–190.

Dieks, D. (1989). Quantum mechanics without the projection postulate and its realistic interpretation. *Foundations of Physics, 38*, 1397–1423.

Dieks, D. (1994). Modal interpretation of quantum mechanics, measurements, and macroscopic behavior. *Physical Review A, 49*, 2290–2300.

Dieks, D. (2007). Probability in modal interpretations of quantum mechanics. *Studies in History and Philosophy of Modern Physics, 19*, 292–310.

Dieks, D., & Vermaas, P. (Eds.). (1998). *The modal interpretation of quantum mechanics*. Dordrecht: Kluwer Academic Publishers.

Dirac, P. A. M. (1958). *The principles of quantum mechanics*. Oxford: Clarendon Press.

Kochen, S. (1985). A new interpretation of quantum mechanics. In P. Mittelstaedt & P. Lahti (Eds.), *Symposium on the foundations of modern physics 1985* (pp. 151–169). Singapore: World Scientific.

Laura, R., & Vanni, L. (2008). Conditional probabilities and collapse in quantum measurements. *International Journal of Theoretical Physics, 47*, 2382–2392.

Lombardi, O., & Castagnino, M. (2008). A modal-Hamiltonian interpretation of quantum mechanics. *Studies in History and Philosophy of Modern Physics, 39*, 380–443.

Lombardi, O., Castagnino, M., & Ardenghi, J. S. (2010). The modal-Hamiltonian interpretation and the Galilean covariance of quantum mechanics. *Studies in History and Philosophy of Modern Physics, 41*, 93–103.

Lombardi, O., & Dieks, D. (2013). Modal interpretations of quantum mechanics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition). http://plato.stanford.edu/archives/fall2013/entries/qm-modal/.

van Fraassen, B. C. (1972). A formal approach to the philosophy of science. In R. Colodny (Ed.), *Paradigms and paradoxes: The philosophical challenge of the quantum domain* (pp. 303–366). Pittsburgh: University of Pittsburgh Press.

van Fraassen, B. C. (1991). *Quantum mechanics: An empiricist view*. Oxford: Clarendon Press.

Vermaas, P., & Dieks, D. (1995). The modal interpretation of quantum mechanics and its generalization to density operators. *Foundations of Physics, 25*, 145–158.

von Neumann J (1932) *Mathematische grundlangen der quantum mechanik*. Berlin: Springer-Verlag. English edition: *Mathematical foundations of quantum mechanics*, 1955. Princeton: Princeton University Press.

Zurek, W. H. (2003). Decoherence, einselection, and the quantum origins of the classical. *Reviews of Modern Physics, 75*, 715–776.

Why I Am Not an Everettian

Foad Dizadji-Bahmani

Abstract Everettian quantum mechanics (EQM) results in “multiple, emergent, branching quasi-classical realities” (Wallace (2009)). The possible outcomes of measurement as per ‘orthodox’ quantum mechanics, are, in EQM, all instantiated. Given this metaphysics, Everettians face the ‘probability problem’ – how to make sense of probabilities and recover the Born Rule. To solve the probability problem, Everettians have derived a quantum representation theorem. There is a notable argument against the soundness of the representation theorem based on so-called ‘branch counting’. Everettians have sought to undercut this argument by claiming that there is no such thing as the number of branches. In what sense is it both true that there is no such thing as the number of branches and that there are multiple branches? Various answers to this question have been given. These can be categorised into two kinds: that there are ‘indeterminately-many’ branches or that there are ‘indeterminably-many’ branches. I argue that neither suffices to undercut the argument against the quantum representation theorem. I conclude that the quantum representation theorem is unsound and that the probability problem facing EQM remains unsolved.

F. Dizadji-Bahmani (✉)
Department of Philosophy, Logic and Scientific Method, London School of Economics
and Political Science, London, UK
e-mail: f.dizadji-bahmani@lse.ac.uk

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_18,

1 Introduction

The measurement problem in quantum mechanics is well-known. The decoherence based Everettian interpretation of quantum mechanics (EQM) solves this problem by dropping the 'collapse postulate'; it purports to provide an account of quantum phenomena based only on unitary evolution.

> The "Everett interpretation of quantum mechanics" is just unitary quantum mechanics, taken literally as a description of the world; it is a "many-worlds" theory because it instantiates multiple, emergent, branching quasiclassical realities. (Wallace 2009, 1)[1]

One of the salient features of EQM is that the possible outcomes of a measurement as per 'orthodox' quantum mechanics, are all instantiated. This ushers in a problem: how to make sense of the probabilities that constitute the empirical content of the original theory? Specifically, if each of the possible outcomes (orthodoxly understood) are instantiated, how can it even make sense to talk of the probability of a particular outcome occurring? This is the *probability problem*. (cf. Wallace 2009; Greaves 2004).

The probability problem can be understood in two parts: first, there is the problem of 'making conceptual room' for probabilities – the *Incoherence Problem*. Second, even if one can make sense of probabilities given this branching picture, how does one recover the probabilities ascribed by the standard Born Rule – the *Quantitative Problem*.

The way Everettians purport to solve the probability problem is to take the 'decision-theoretic approach'.[2] The decision-theoretic approach: construe quantum probabilities as the subjective degrees of beliefs of rational agents facing branching situations; operationalise those subjective degrees of belief to the betting preferences of such agents to hypothetical quantum games (games where the agents future selves receive payouts depending on the outcomes of branching); from a set of decision theoretic axioms, which constrain the betting preferences of rational agents partaking in the hypothetical quantum games, prove a representation theorem to the effect that rational agent's degrees of beliefs in the outcomes of measurement events must numerically coincide with the standard Born Rule.[3]

How can the use of decision theory be motivated in this context? After all, decision theory is a normative and idealised framework for dealing with decision making situations, situations where there are various *possible* outcomes with associated consequences. Yet in the Everettian universe the possible outcomes as

[1] I do not consider the older "Many-Worlds" or "Many-Minds" interpretations, as these are no longer pursued.

[2] This was first put forward by Deutsch (1999) and has been greatly developed by Simon Saunders and David Wallace (2008). However, it should be noted that not all Everettians consider there to be a probability problem.

[3] For details see Wallace (2009).

per ‘orthodox’ quantum mechanics *are all realised*. Everettians have suggested two ways in which to motivate the decision theoretic approach:

- “The *subjective uncertainty (SU) viewpoint*: Given that what it is to have a future self is to be appropriately related to a certain future person, and that in normal circumstances I expect to become my future self, so also in Everettian splittings I should expect to become one of my future selves. If there is more than one of them I should be uncertain as to which I will become; furthermore, this SU is compatible with my total knowledge of the wavefunction and its dynamics.” (Wallace 2007, 313)
- “The *objective-determinism (OD) viewpoint*: Branching leads, deterministically, to my having multiple future descendants. Rationally speaking, I should act to benefit my future descendants, for exactly the same reason that people in non-branching possible worlds would act to benefit their single descendant. Situations of conflict may arise between the interests of my descendants (such as when I bet on one possible outcome of a measurement), in which case I will have to weigh up how much I wish to priorities each descendants interests.” (ibid)

SU is the position advocated by Wallace (2003, 2007, 2009, 2010) and Saunders (1998). OD is explicitly defended by Greaves (2004). Let us assume for the sake of exposition, that each of these approaches suffices to motive the decision theoretic turn. That is, that each suffices to make sense of using decision-theory to operationalise the betting preferences of rational agents faced with branching. Successfully motivating the use of decision theory in this way solves the Incoherence Problem. What about the Quantitative Problem? This is solved by deriving the representation theorem.

EQM seems to be an attractive interpretation of quantum mechanics. It avoids the measurement problem, (and as such) is not demonstrably not universally applicable. Contrary to naysayers it is ontologically parsimonious; it’s only ontological posit is the quantum state – all other objects are higher-order patterns in this fundamental ontology.[4] It is deterministic but can account for the stochasticity of quantum phenomena, and, finally, it is Lorentz invariant. I want to believe in this interpretation. Unfortunately, I do not think that EQM is tenable, as it stands. What I argue in this paper is that the purported solution to the probability problem fails on closer inspection.

2 The Probability Problem

Suppose that the SU approach is correct for the sake of exposition. Given that under SU a rational agent is to think that she will become one of her post-branching-selves but uncertain as to which one that is, what is the probability, understood as

[4] I subscribe to the view that ontological parsimony is fixed, *ceteris paribus*, by the number of *kinds* of objects not by the number of objects.

her subjective degree of belief, of seeing one outcome in particular? (There is, of course, a corresponding OD question.)

What the quantum representation theorem requires is that, on pain of irrationality, she considers the probability of the outcome as the ‘weight’ of the branches in which those particular outcomes occur. Let us bracket any worries about the metaphysics of the ‘weight’ of branches for now. The important point is that the ‘weight’ of the branches numerically coincides with the standard probabilities as per the Born rule. That is, the quantum representation theorem purportedly shows that a rational agent must set her credence according to the Born rule in the face of Everettian branching.

The representation theorem is derived from a set of axioms, which come in two classes: *axioms of richness* and *axioms of rationality*, and it is with the latter I am presently concerned. That she is rationally required to set her degrees of belief so as to coincide with the Born rule is precisely because the (latter) axioms are dictates of rationality when faced with branching, or at least Everettians contends.

In order to be satisfied by all this then, what needs to be shown is that the *axioms of rationality* are indeed rational. In his most recent paper, Wallace presents ten axioms from which he derives the quantum representation theorem. The details do not matter here, suffice to say that none of them are redundant.[5] Consider one of these, the ‘Branching Indifference’ axiom:

- *Branching indifference* (BI): “An agent is rationally compelled to be indifferent about processes whose only consequence is to cause the world to branch, with no rewards or punishments being given to any of his descendants.” (Wallace 2007, 327)

Notice that BI is not saying that you need not care about the number of your future descendants. BI is stronger than that: it requires you *not to care*. Is this rational? Wallace himself worries: it is “not at all obvious [that BI is rational]: why should I not care about whether there is one of me or a 1010 min from now?” (Wallace 2007, 327) Quite.

So what arguments can be produced in it’s favour? Wallace offers two. The first argument is only available under SU, and runs as follows. Suppose that the number of branches is epistemically inaccessible. The scenario where there are a 100 descendants each either seeing outcome A exclusive-or outcome B is epistemically equivalent to the scenario where there is just one descendant seeing each kind of outcome. But under SU, claims Wallace, only the distribution of the kinds of outcomes is relevant to forming your degrees of belief about what you will see. Thus, under SU you will be indifferent between these epistemically equivalent scenarios, and hence, BI is rational. I only sketch this argument for I think that

[5] Here I consider the proof given in Wallace (2009) as it is the most recent. Earlier proofs (Wallace, 2003, 2007) have this axiom (or one relevantly similar) and my arguments here carry over straightforwardly.

Greaves (2004) has successfully argued that SU is untenable. But there is a further reason that this is unpersuasive. As aforementioned, the argument (irrespective of its force) is available under SU; there is no positive argument for BI under OD. And this is troubling because, as Wallace himself says, "[i]t is almost impossible *not* to accept the OD viewpoint as valid, since it is just a literal reading of the physics." (Wallace 2007, 314). So, here is a putative axiom of rationality – BI – for how to constrain one's degrees of belief when facing branching for which there is no positive argument under a literal reading of the physics. The first argument for BI is not persuasive.

The second argument that Wallace offers in defence of BI is that it is the only available strategy: "it is not in fact possible to pursue a non-branch-indifferent strategy in a quantum universe." (Wallace 2007, 314), he writes. That is, BI is the *unique* strategy, or so it is claimed.

Grant for a moment that BI is the unique strategy. Even if so, it does not follow that it is a *rational* strategy. From a purely logical point of view, one needs a further premise: that there is at least one rational strategy in an Everettian universe.[6] To reach the desired conclusion, viz. that BI is the *unique rational* strategy this further premise needs motivating. This has not been done, and it is hard to see how it could be.

A different (i.e. non-BI) and very suggestive way for an agent to set her degrees of belief for seeing a particular outcome is to the ratio of the number of her future descendants who see that kind of outcome to the total number of descendants – call this the Branch Counting strategy (BC). BC certainly seems rational. After all, she stands in the same relation of personal identity over time to each of her descendants which, under SU, suggests that she should regard herself as equally likely to become any of them in particular, and which, under OD, suggests that she should value each of their lives equally. And this is precisely what BC encodes.

Clearly if BC is rational then this undermines the claim that BI is the unique rational strategy, and thereby the representation theorem. And that BC *is* a rational strategy has not been questioned per se. Rather, Wallace has argued that it is not possible to pursue BC, and that this entails it is not a rational strategy, and, hence, that it (BC) does not undermine the representation theorem. In this paper I do not defend BC as the *unique rational strategy*. I merely assume that, modulo defeating reasons such as the ones Wallace proffers, BC is one rational way to set one's credences in the face of branching.

[6] One may be tempted to say that if BI is literally the only available strategy then it must be rational to pursue it – surely it is rational to pursue the only available strategy? But here push comes to shove: it is not the case that BI is literally the only strategy – one could set one's degrees of belief, say, at whim. (And notice that whim might, *de facto* violate BI!) In any case, the salient point is that rationality does not come by default. With regards, say, whim, one would need positive reasons to think it rational to do so, *and so too with BI*.

2.1 What's Wrong with BC?

What is wrong with branch counting? Here is what Wallace says:

> [B]ranch counting [cannot] actually be motivated or even defined given the structure of quantum mechanics. There is no such thing as "branch count"... the branching structure emergent from unitary quantum mechanics does not provide us with a well-defined notion of how many branches there are. All quantum mechanics really allows us to say is that there are *some* versions of me for each outcome... (Wallace 2009, 28 (emph. added))

It seems contradictory to say both that there is no such thing as the number of branches and that there are multiple branches – what can 'multiple' mean other than 'an unspecified number of' here? And what can the weights, by which one is to set one's credences, be weights *of*, if not sets of branches? But before turning to these questions, in Sect. 3, consider the following.

Within the decision theoretic framework the branch count is well-defined. (cf. Wallace 2009, 28) So even if it is true that there is no such thing as the number of branches 'really', the above does not constitute a persuasive argument against BC. After all, the representation theorem is proved *within* just this decision theoretic framework and its soundness depends on the veracity of the framework. Coarsely put, if framework used to derive the quantum representation theorem is good enough, then one cannot deem BC to be in some sense, illegitimate.

In response Wallace suggests that a way to understand the axioms from which the quantum representation theorem is derived is as: "prohibitions on strategies [like BC] that just exploit artefacts of [the framework]." (Wallace 2010, 13) By Wallace's lights, whilst the decision-theoretic framework is such that it allows for the formation of the alternative strategies, like BC, his axioms indicate which alternative strategies (would) 'exploit artefacts' of the framework. In particular, the artefact of numerous individuated branches. But Wallace's response is not compelling: it's hard to see why his axioms do not 'exploit artefacts' of the model in just the same way as alternative strategies purportedly do. This is best seen by an example: consider the "State Supervenience" axiom which requires that an agent's preferences between acts "depend only on what physical state they actually leave his branch in..." (Wallace 2009, 12) This axiom is only tenable in so far as there is a referent to 'his branch' but this requires individuated branches – some future branch in which the agent finds himself. But, we are told that there is no such thing as individuated branches 'really', that individuated branches are a mere artefact of the theory.

The dilemma for the Everettian is this: if the notion of numerous individuated branches is an untenable idealisation, then the quantum representation theorem is unsound for the axioms from which it is derived rely on this idealisation too. On the other hand, if the idealisation that the framework involves is tenable, it remains to be shown that BC is not rational.

Which ways out of this dilemma for the Everettian? To get out via the second horn would require showing that *even within* the idealised framework branch counting is not a rational strategy. Some such demonstration is not forthcoming.

To get out via the first horn would require an argument that the axioms used do not after all rely on the idealisation. With regards the ‘State Supervenience’ axioms this would require showing that it is tenable even if there is no such thing as individuated branches, for example. In essence, this comes down to the issue about there being ‘no such thing’ as the number of branches yet there being some – whether this can be understood in such a way as to make the axioms used to derive the representation tenable whilst undercutting BC. In Sect. 3, I show that this way out of the dilemma is not feasible.

3 Number of Branches

In what sense is it both true that “there is no such thing as the number of branches” and that there are “multiple branches”? Doesn’t the claim that there are multiple branches commit one to there being a number of them? The claim there this no such thing as the number of branches is crucial in EQM for it serves to undermine BC.

Wallace offers various arguments for there being no such thing as the number of branches, which entails that branch counting is not available. But none are very clear. In particular, Wallace equivocates between metaphysical and epistemological reasons for why one cannot count the number of branches. One can read him as claiming there just is literally no such thing as the number of branches – this is the metaphysical sense. On the other hand, he could be understood as making an epistemological claim: branching is so complicated as to preclude any possibility of counting the branches. In what follows, I consider each of these readings in turn and show that neither satisfactorily undermines BC strategy.

Let me introduce the following locutions for the metaphysical and epistemological readings respectively. Metaphysical: The number of branches is *indeterminate*. Epsitemological: The number of branches is *indeterminable*. I now argue that neither reading affords a resolution to the dilemma facing the Everettian and as such the probability problem remains unsolved.

3.1 The Number of Branches Is Indeterminate

One locution for the claim that there is no such thing as the number of branches is that the number is *indeterminate*. As Wallace writes: “All quantum mechanics really allows us to say is that there are *some* versions of me for each [kind of] outcome” (Wallace 2009, 28). On this reading, whilst there are some branches there is no fact as to the number of them, *the number is indeterminate*. Crucially, we are told that this fact is underpinned by decoherence, which gives rise to the branching structure.

Does decoherence support the claim that the number of branches is indeterminate? The formalism of quantum mechanics is such that there can be superpositions of states. If we take this formalism to be a literal description of microscopic objects, then these micro-objects are ascribed indefinite properties. This is counter-intuitive given a classical world-view, but it is also pathological when it comes to macroscopic objects. Still, if the macro-objects are just composed of micro-objects the linearity of the formalism requires that, *pace* experience, the macro-objects do not have definite properties, that they are not in definite states. The leading idea of the decoherence-based Everettian approach is that "*a pattern view of macroscopic ontology essentially solves the problem of indefiniteness by replacing indefiniteness with multiplicity.*" (Wallace 2003, 98)

Consider the example that is given to illustrate multiplicity replacing indefiniteness, namely the Schrödinger's Cat thought-experiment. We imagine a cat sealed into a box with a device primed to measure an unstable atomic nucleus at 12 noon, such that a poison is or is not released into the depending on the outcome. Suppose we set this up at 11 a.m. and consider what will happen in the next hour.

> If the atom's state is indefinite just before the measurement, then so is the cat's state just after the measurement… Now, consider the evolution of the system after 12 noon, when the measurement is made, but suppose that the atomic nucleus, instead of being in an indefinite state, either definitely did or definitely did not decay… [I]f the cylinder of poison gas breaks, then cat psychology tells us that the cat will probably jump backwards, and animal physiology tells us that it will die and in due course start to decompose. Now, quantum mechanics is linear. If we know what happens if the atom definitely does, or definitely does not, decay, then we can predict what happens if we have a superposition of decaying and not decaying. However, in doing so we are using exactly the same methods as before: we are taking advantage of the patterns present in the two branches of the wave-function. In other words – and this is the crucial point – in each of the branches there is a 'cat' pattern, whose salience as a real thing is secured by its crucial explanatory and predictive role. Therefore… there is a cat present in both branches after measurement. (Wallace 2003, 93)

What role does decoherence play? A crucial one for "*without it, we would not have the sort of branching structure which allows for the existence of effectively non-interacting multiple near-copies of a given process.*" (ibid. 102) That is, what decoherence (putatively) ensures is that a small system (like the atomic nucleus) when entangled with a larger environment (like the measurement device and the rest of the box) evolves such that very quickly there emerge almost (but not entirely) non-interfering branches each of which macroscopically correspond to either alive or dead cats.

It is important to be clear that this is an interpretation of decoherence models. Formally, the models show that the value of the coefficients of the cross-diagonal terms in a certain basis rapidly diminish and the sum of the values of coefficients of the diagonal terms tends to one. The cross-diagonal terms are taken to represent the interference between the various 'branches', where the 'branches' are represented by the diagonal terms which in this basis correspond to the phenomenologically-correct macroscopic outcomes. That the value of the coefficients of the cross-diagonal terms rapidly diminish is interpreted as there emerging effectively non-interacting branches.

Now, in the given example, Wallace has it that decoherence leads to only two branches. This can be understood as a pedagogical simplification: there may well be many dead cats and many alive cats, in corresponding numerically distinct though qualitatively identical branches. But, and this is the crucial point, the number of branches is not *indeterminate*. The cornerstone of the decoherence-based Everett interpretation *just is* that decoherence gives rise to the emergence of multiple patterns in the fundamental quantum state.

There are time-scales at which there *is no number of* cats, people, etc. for the simply reason that these patterns are yet to emerge in the quantum state:

> Before the decay there is certainly one cat. When the measurement occurs we will have a coherent superposition of both measurement outcomes but after a very short time decoherence will remove the interference between these branches, and after this time there will be two cats present. (Wallace 2003, 97)

Again, this may be a simplification in the sense that there may be more than two cats (two branches) but just how many there are is not something which can be indeterminate, at least if the interpretation of the decoherence models that Everettians advocate is correct. Under this interpretation patterns emerge in the wavefunction corresponding to live and dead cats, and cats are not the kind of things of which there can be an indeterminate number of. No amount of hand-waving about just how very complicated this process is will support the claim that it is both the case that decoherence leads to some branches but that there is no such thing as the number of branches, understood metaphysically. Finally it is worth stressing that there must be, at least something like, individuated branches: what are the weights to which Everettians make reference, weights *of*, otherwise?

3.2 The Number of Branches Is Indeterminable

In other passages, Wallace seems to want to undercut BC by arguing that the number of branches is *indeterminable*. So, for example, we are told that "[r]ealistic models of macroscopic systems are invariably infinite-dimensional, ruling out any possibility of counting the number of discrete descendants." (Wallace 2007, 328)

What are we to make of the indeterminability claim? In general, in any actual decision making situation it may not be possible to decide on a course of action as per some decision theoretic strategy because of one's epistemic limitations. However, that the contingencies of a situation are such that one cannot abide by the strategy does not show that the strategy is not rational. The rational strategy is quite often only hypothetically implementable. Moreover notice that for any realistic physical system we do not have epistemic access to the weights of outcomes either. Indeed the Born rule is also only hypothetically implementable.[7] Merely pointing out that BC is not practically implementable does not suffice to secure BI as the unique rational strategy.

[7] Furthermore, the decoherence models that underpin the Everettian interpretation are highly idealised.

4 Conclusion

The purported solution to the probability problem is to take the 'decision-theoretic turn': construe quantum probabilities as the degrees of belief of idealised rational agents facing branching; operationalise those degrees of belief to their betting preferences in quantum games; and derive a quantum representation theorem recovering the Born Rule via axioms of rationality that constrain those preferences. The promises of EQM are enticing and this purported solution to the probability problem is both ingenious and formally elegant.

However, the solution is not successful. The Everettian faces a dilemma: either the derivation of the quantum representation theorem relies on an untenable idealised framework or at least one of the axioms used to derive is not an axiom of rationality. I considered the ways out of this dilemma and showed them to be wanting. I conclude that probability problem in EQM is not solved.

Acknowledgements I am grateful to the following people for interesting and instructive discussions/comments pertaining to this paper: Jason McKenzie Alexander, Seamus Bradley, Nazim Bouatta, Erik Curiel, Roman Frigg, Conrad Heilmann, Eleanor Knox, John Norton, Matt Parker, Wolfgang Pietsch, Miklós Rédei, Marie Toseland, Alastair Wilson, and members of the audience at EPSA 2011, Athens. I am particular grateful to comments I received from an anonymous referee; I found the comments insightful and they prompted several important revisions.

References

Deutsch, D. (1999). Quantum theory of probability and decisions. *Proceedings of the Royal Society of London A, 455*, 3129–3137.

Greaves, H. (2004). Understanding Deutsch's probability in a deterministic multiverse. *Studies in History and Philosophy of Modern Physics, 35*(3), 423–456.

Saunders, S. (1998). Time, quantum mechanics, and probability. *Synthese, 114*, 373–404.

Saunders, S., & Wallace, D. (2008). Branching and uncertainty. *British Journal for the Philosophy of Science, 59*, 293–305.

Wallace, D. (2003). Everett and structure. *Studies in the History and Philosophy of Modern Physics, 34*, 86–105.

Wallace, D. (2007). Quantum probability from subjective likelihood: Improving on Deutschs proof of the probability rule. *Studies in History and Philosophy of Modern Physics, 38*, 311–332.

Wallace, D. (2009). A formal proof of the Born rule from decision-theoretic assumptions. ArXiv: 0906.2718v1. Accessed on 11 November 2013.

Wallace, D. (2010). Decoherence and ontology (or: How I learned to stop worrying and love FAPP). In S. Saunders, J. Barrett, A. Kent, & D. Wallace (Eds.), *Many worlds? Everett, quantum theory, and reality*. Oxford: OUP.

The Emergence of Integrability in Gauge Theories

Nazim Bouatta and Jeremy Butterfield

Abstract The aim of this paper is to contribute to a better understanding of quantum field theories by discussing a famous physical limit, the ’t Hooft limit, in which the theory concerned often simplifies. This limit is important in much current research connecting quantum field theory with string theory (and thus with gravity). The idea of the limit is that the number N of colours (or charges) goes to infinity. The simplifications that can happen in this limit, and that we will consider, are: (i) the theory’s Feynman diagrams become planar, and (ii) the theory becomes integrable, and indeed corresponds to a well-studied system, viz. a spin chain. We see this as a case in which one theory is emergent from, yet reduced to, another one.

1 Introduction

The aim of this paper is to analyse philosophically the ’t Hooft limit in quantum chromodynamics (QCD) and other gauge theories, in which the number N of colours (or charges) goes to infinity. The limit is controlled by requiring that the product $\lambda := g^2N$ (of the square of the coupling constant g with N) is fixed. λ is called ‘the ’t Hooft coupling’. We will connect the technicalities of this limit to the idea that in such a limit, one theory (or a sector, or regime, of a theory) can be emergent from, and yet reduced to, another one.

N. Bouatta (✉)
Darwin College and DAMTP, Cambridge, CB3 0WA, UK
e-mail: nb379@cam.ac.uk; nb379@hermes.cam.ac.uk

J. Butterfield
Trinity College, Cambridge, CB2 1TQ, UK
e-mail: jb56@cam.ac.uk; jb56@hermes.cam.ac.uk

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_19,

The ’t Hooft limit is a rich subject, with many aspects which we cannot pursue here. One main one, which we will study in a companion paper, is that this limit sheds light on the connection between quantum field theories and gravity—a connection much studied under the label ‘AdS/CFT’. Here we will instead be concerned with emergent planarity, mainly in QCD (Sect. 2), and emergent integrability, mainly in a supersymmetric cousin of QCD (Sects. 3 and 4). The main idea will be that one can prove that in the ’t Hooft limit some quantum field theories are integrable: and furthermore, correspond to certain well-studied systems, viz. some spin chains.

Thus we will first sketch, in Sect. 2, some relevant aspects of gauge theories, especially QCD. The main point will be a surprising simplification. QCD is described by a gauge group $SU(3)$, where 3 is the number of colours: the gauge field is described by a 3×3 matrix. So one naturally expects that theories using a value of N higher than 3, i.e. using a larger matrix, will be increasingly complex. But in 1974, ’t Hooft discovered that surprisingly, the theory simplifies at $N = \infty$. In the perturbation expansion, only those Feynman diagrams that can be drawn on a plane (‘planar diagrams’) remain. This first main topic is associated with the appearance of string-like structures: which leads to the topic above, AdS/CFT.

In Sect. 3, we describe the novel features and mathematical structures that can occur in the $N = \infty$ limit. But we ‘change horses’, i.e. we consider a different theory. For QCD remains in this limit a complicated theory. So we emphasize, as the physics literature does, a simpler theory, again in this limit: maximally supersymmetric Yang-Mills theory. Despite the name, this is simpler than QCD! It is also guessed and hoped to be one of the essential models for understanding more complicated theories. So much so that nowadays it is sometimes dubbed ‘the harmonic oscillator of quantum field theory’. (We emphasize that our discussion of this theory does not require that nature actually be supersymmetric.)

In Sect. 4, we arrive at our second main topic: integrability at the $N = \infty$ limit. That is: we discuss how, thanks to planarity, certain physical aspects of these field theories are mapped into integrable spin chains. Here we connect with an old dream of quantum field theory: to calculate analytically the mass-spectrum of a theory, i.e. the masses of its particles such as a proton, as a function of the parameters of the theory, such as coupling constants and the energy-scale. The theory at issue here, maximally supersymmetric Yang-Mills theory, is conformally invariant, i.e. invariant under a transformation that changes scales but preserves angle. This means the theory has no massive particles. But there is an analogue of the mass that we can aspire to compute: namely, the scaling dimension of local operators. The idea is that the correlation function of such an operator, i.e. the correlation of its values at different spacetime points (as given by an expectation value of a product), falls off with some power Δ of the spatiotemporal distance between the points concerned. This Δ is the scaling dimension: it includes a quantum addition (called ‘anomalous dimension’) to the term Δ_0 obtained by classical dimensional analysis (the ‘bare dimension'). And it has recently been shown that Δ can be calculated by analysing the associated spin chain systems: a special case of the old dream.[1]

Two *caveats* about rigour. (1) Various results that we assume, e.g. that maximally supersymmetric Yang-Mills theory is conformally invariant, and the recent results on integrability that we focus on, are *not* proven rigorously as one would hope for, and as in e.g. theorems in AQFT. For example, these results are obtained by computations in an expansion. (2) In $3 + 1$ dimensions, integrability of a finite N interacting quantum field theory is surely rare indeed. But notwithstanding (1) and (2), we submit that these recent results are such a significant advance towards a more rigorous understanding of interacting quantum field theories, that it is worthwhile for philosophers to "dive in".

2 The 't Hooft Limit in QCD

QCD is a rich and complicated theory. Several of its essential features are still poorly understood, within physics—let alone philosophy! These include confinement, dynamical mass generation and chiral symmetry breaking. Besides, confinement means that at low energies, QCD becomes strongly coupled; and accordingly, increasingly difficult to calculate. (This is the 'flip-side' of the fact that QCD is asymptotically free, i.e. that the effective coupling constant decreases as the energy increases.) The best available approach to deal with this is to use numerical simulation on a lattice.

But even apart from calculating specific problems, QCD is so complicated a theory that we cannot expect to obtain exact solutions. Therefore, we need to find some sort or sorts of approximation scheme. Since a good scheme is traditionally considered to require an appropriate expansion parameter, we face the question: what possible expansion parameter does QCD contain?

The ordinary coupling constant is not really a free parameter in QCD, because as a result of the renormalisation group flow, it is absorbed into defining the scale of masses by dimensional transmutation. Indeed, this is one of the most important facts we know about QCD: not least because this is what makes the theory both difficult to calculate and hard to understand. Thus the theory has no obvious free parameter that could be used as an expansion parameter: it is apparently a theory without parameters. Thus one must hope to find a non-obvious free parameter.

Famously, 't Hooft (1974) pointed out that QCD has a non-obvious candidate for an expansion parameter. Namely, he suggested that one should generalize QCD, from three colours and an $SU(3)$ gauge group, to $N \equiv N_c$ colours and a $SU(N)$

[1] This hope is called the 'old dream' at the start of the excellent recent review by Beisert et al. (2012). We stress that these recent advances arose from examining the connection between conformal field theories and gravity.

gauge group.[2] More precisely, he expanded the partition function and the correlation functions of a $SU(N)$ gauge theory in powers of N; and argued that the theory simplifies when the number N of colours is large. Complicated Feynman diagrams at finite N are replaced by much simpler *planar* diagrams, i.e. diagrams that can be drawn on a plane, and so one has to cope with far fewer diagrams. In a bit more detail: using a new kind of diagram (with double lines to represent a colour index contraction), 't Hooft showed that at large N, diagrams that cannot be drawn on a plane will be suppressed, compared with those that can, by a factor that is a power of N. And since the dominant diagrams at large N can be drawn on a plane, they look like two-dimensional surfaces. This prompted the idea that these surfaces could be analysed as the propagation in time of a one-dimensional object, i.e. a string. Thus planarity, in addition to simplifying the theory, suggested a connection between quantum fields and strings.

Besides, the dominance of planar diagrams has proven very useful in lattice QCD (cf. Teper (2009) for a recent review); and has prompted the hope that one could solve the theory exactly at $N = \infty$, and then one could better understand QCD itself by doing an expansion in $1/N = 1/3$. Although these hopes have not yet come true, the results to be surveyed in the rest of this paper are surely progress.

3 At the $N = \infty$ Limit of Quantum Field Theory

3.1 *Introducing $\mathcal{N} = 4$ Super Yang-Mills Theory*

We now 'change horses', i.e. consider a different theory than QCD or $SU(N)$. Namely: $\mathcal{N} = 4$ maximally supersymmetric Yang-Mills theory, the study of which was initiated by Brink et al. (1977); for short: '$\mathcal{N} = 4$ SYM', or even just 'SYM'. Here, $\mathcal{N}$ is the number of copies of the supersymmetry algebra: *not* the number of colours, N. However, the theory's gauge group will be the familiar $SU(N)$.

We mentioned in Sect. 1 that this theory's $N = \infty$ limit is simpler to study than that of QCD, and also exhibits planarity and integrability (details in Sect. 4). But there are also two other good reasons to study it.

First, it has various remarkable properties. It has a large amount of symmetry: which is the origin of the planarity and integrability just mentioned. More specifically, it is conformally invariant, implying that it has no inherent scale. Classically, many theories are conformal, e.g. theories with only massless fields; (of course, Maxwell theory is the paradigm example). But $\mathcal{N} = 4$ SYM stays conformal

[2] For an introduction, we recommend Coleman (1985). 't Hooft's proposal had precedents in 1960s work in statistical mechanics; cf. Brezin and Wadia (1993).

even at the quantum level. In particular, its β-function (which describes how the coupling constant depends on the energy scale) is believed to be zero to all orders in perturbation theory. And although QCD is not conformal, its being asymptotically free means that at high energies it is close to being conformal.[3] Thus many essential features of high energy gluon scattering—which is very relevant for the LHC—can be analysed by studying gauge boson amplitudes in $\mathcal{N} = 4$ SYM.

Second, this theory is the gauge theory 'side' of the best-understood example of the gravity/gauge, or AdS/CFT, correspondence. (The gravity side is a certain string theory on a cousin of anti-de Sitter space: hence the label, with 'CFT' standing for 'conformal field theory'—here $\mathcal{N} = 4$ SYM.) In Sect. 1, we postponed this topic to another paper. So suffice it to say here that since Maldacena (1997) introduced this correspondence, it has produced stunning insights on both sides. Indeed, the significance of this paper's main topic, integrability, lies largely in the light it sheds on AdS/CFT.

In Sect. 3.2, we will first discuss the dilatation operator of $\mathcal{N} = 4$ SYM, and anomalous dimensions. Then we introduce the *single trace operators*: these function as 'building-blocks' of the gauge invariant operators, in the large N limit. We end the section by introducing the relation to spin chains, which looks forward to Sect. 4's discussion of integrability.

3.2 Dilatation, and Single Trace Operators

The generator of dilatations D turns out to play a crucial role in the *quantum* structure of $\mathcal{N} = 4$ SYM. While the generators of the Poincaré subgroup of the theory's overall symmetry group (which is in fact: $PSU(2, 2|4)$) do not get any quantum corrections, the dilatation operator D *does*—notwithstanding the theory's being conformally invariant. Thus:

$$D = D_0 + \Delta D(g)\,, \tag{1}$$

where D_0 is the classical operator and ΔD is the anomalous dilatation operator which depends on the gauge coupling g.

Now let $\mathcal{O}(x)$ be a local operator in the field theory with scaling dimension Δ. The physical idea of Δ is that it is the analogue of the mass and the mass

[3] More precisely: classical QCD *is* conformally invariant, but we have every reason to believe that when it is quantized, the conformal symmetry is broken, with only Poincaré symmetry remaining. Proving this is one of the famous Clay Millennium prizes. The general topic of classical symmetries being lost after quantization ('anomalies') suggests in QCD an approach to the origin of mass quite different from the usual Higgs mechanism's postulation of an additional field.

spectrum in QCD. Technically: under the rescaling $x \to \lambda x$, the operator $\mathcal{O}(x)$ scales as $\mathcal{O}(x) \to \lambda^{-\Delta}\mathcal{O}(\lambda x)$; and the dilatation operator D is the generator of these scalings, by which we mean that $\mathcal{O}(x) \to \lambda^{-iD}\mathcal{O}(x)\lambda^{iD}$. The dimension Δ is $\Delta_0 + \gamma$; with Δ_0 the bare dimension corresponding to the classical operator D_0 in Eq. (1) and γ the anomalous dimension arising from quantum corrections corresponding to ΔD. Thus to find the anomalous dimension of $\mathcal{O}(x)$, one considers its two-point correlator with itself:

$$\langle \mathcal{O}(x)\overline{\mathcal{O}}(y)\rangle \approx \frac{1}{|x-y|^{2\Delta}}\,. \tag{2}$$

We now turn to the operators one actually encounters in $\mathcal{N} = 4$ SYM. The punchline, for the rest of the paper, will be that the anomalous dimension of a kind of operator, called *single trace operators* will, for large N, be encoded in the Hamiltonian of a spin chain. (In fact, each site in the chain carries a representation of $SO(6)$, which is a subgroup implementing supersymmetry (called: the R-symmetry subgroup) of the theory's overall symmetry group, $PSU(2, 2|4)$.)

We recall that the physical observables of a gauge theory are gauge invariant operators. In $\mathcal{N} = 4$ SYM, the local gauge invariant operators are made up of products of traces of the fields that transform covariantly under the gauge group $SU(N)$. These fields include the scalars ϕ (of which there are six), the fermions ψ (of which there are eight), the field strengths of the gauge fields, and all these fields' covariant derivatives. It is thus clear that the single trace local operator

$$\mathcal{O}(x) = \mathrm{Tr}[\chi_1(x)\chi_2(x)\ldots\chi_L(x)]\,, \tag{3}$$

where the trace is over the internal degree of freedom indices, and $\chi_i(x)$ is one of the above covariant fields (with or without covariant derivatives), is itself gauge invariant. We can also build other local gauge invariant operators by taking products of traces.

In Sect. 4, we will take the 't Hooft limit, where the number of colours N is large. This limit has the remarkable property that the scaling dimension of the product of single trace operators is equal to the sum of their scaling dimensions, so that all information about the spectrum of local operators is determined by the single trace operators. Thus, for computing dimensions in this limit, it suffices to concentrate on single trace operators.

We can already state the key idea about how all this relates to spin chains. For a single trace operator with L arguments, $\mathcal{O}(x) = \mathrm{Tr}[\chi_1(x)\chi_2(x)\ldots\chi_L(x)]$, we consider a spin chain of length L, each of whose sites carries a representation of the R-symmetry group $SO(6)$. Thus each site corresponds to one of $\mathcal{O}$'s arguments. And this correspondence is very informative. In particular: the anomalous dimensions of single trace local operators will, for large N, be encoded in the Hamiltonian of the corresponding spin chain.

4 Integrability at the Limit and the Relation to Spin Chains

In Sect. 4.1, we will first outline the computation of anomalous dimensions: recall Sect. 1's old dream of computing a quantum field theory's mass spectrum, and Sect. 3.2's introduction to anomalous dimensions. More precisely: we discuss the one-loop anomalous dimensions for a general set of single trace operators; (recall that in the large N limit the single trace operators encode all the spectral information). We will see how in the large N limit, the contributions to the anomalous dimensions are dominated by diagrams that can be drawn on a plane. Then in Sect. 4.2, we describe the mapping of the system into the problem of computing the energies of a certain spin chain. That is: the one-loop anomalous dimensions will be given by the eigenvalues of the corresponding spin chain's Hamiltonian.

4.1 Computing Anomalous Dimensions

In this subsection we will concentrate on the one-loop anomalous dimensions for single trace operators composed of scalar fields ϕ with no covariant derivatives. Recall that the anomalous dimension is given by the exponent in the two-point correlator of the operator with itself. All scalar fields have bare (classical) dimension 1; and so for single trace operators made up only of scalar fields with no covariant derivative, the bare dimension of the operator is L, the number of arguments i.e. scalar fields inside the trace.

If the coupling constant g is small, then the anomalous dimension γ is much smaller than the bare dimension Δ_0: $\gamma \ll \Delta_0$. In this case we can approximate the correlator in Eq. (2) as

$$\langle \mathcal{O}(x)\overline{\mathcal{O}}(y)\rangle \approx \frac{1}{|x-y|^{2\Delta_0}}(1-\gamma\ \ln \Lambda^2|x-y|^2)\,, \tag{4}$$

where Λ is the cutoff scale. The leading, i.e. classical, contribution to this correlator, $1/|x-y|^{2\Delta_0}$, is called the 'tree-level contribution'.

Let us now summarize what happens as we let $N \to \infty$. There are two main points.

1. The ideas in Sect. 2, about diagrams that can be drawn on a plane coming to dominate the expansion, apply here also. More precisely: nonplanar diagrams will be suppressed by powers like $1/N^2$, where the power depends on the topology of the diagram.
2. The anomalous dimension γ is encoded in an operator Γ, whose eigenvalues are γ: which in Sect. 4.2 will be the Hamiltonian (with nearest-neighbour interactions) of a spin chain.

4.2 Spin Chains

Now we are ready for the punchline. In a very impressive collective effort,[4] it has been shown that the entire class of scalar single trace operators of length L can be mapped to the Hilbert space of a spin chain, i.e. a tensor power of a *finite*-dimensional Hilbert space

$$\mathcal{H}_1 \otimes \mathcal{H}_2 \cdots \otimes \mathcal{H}_\ell \otimes \cdots \otimes \mathcal{H}_L \,; \quad \mathcal{H}_\ell \cong \mathcal{H}_{\ell'} \,. \tag{5}$$

Here each $\mathcal{H}_\ell$ is the Hilbert space for an $SO(6)$ vector representation. In other words: the Hilbert space is that of a one-dimensional spin chain with L sites, where at each site there is an $SO(6)$ vector "spin". It also turns out that we can treat Γ as a Hamiltonian (with only nearest-neighbour interactions) on the spin chain. The energy eigenvalues then correspond to the possible anomalous dimensions for the scalar operators. Although we will not show it here, the Hamiltonian that corresponds to Γ for the spin chain is integrable. And it means that the system is solvable, at least in principle.

We should add that our discussion has been confined to one loop calculations. Going beyond one loop, one finds that the n-loop contribution to the anomalous dimension can involve up to n neighbouring fields in an effective Hamiltonian. Therefore, as N increases, these longer-range interactions become more important; so that at strong coupling the spin-chain is effectively long-range. In this case, the Hamiltonian is not known above the first few loop orders.

After this technical summary, we will briefly discuss integrability's significance. As we announced in Sect. 3.1: its main significance is to shed light on the conjectured AdS/CFT correspondence. Indeed, although our exposition has not stressed the fact: most of the results reviewed above have used, or been inspired by, string-theoretic ideas and results; and often, ideas and results about AdS/CFT.

Though many questions remain open, there is reasonable hope that these integrability results will teach us how to go back to the physically relevant case of QCD, and finally arrive at the long-sought dual description of it by a string theory. It may even take us closer to realizing the quantum field theorist's ultimate goal, unfulfilled for more than 80 years: completely understanding an interacting relativistic quantum field theory in the four space-time dimensions that we are familiar with.

[4] For a review we recommend Beisert et al. (2012) and Beisert (2005).

5 Conclusion

Finally, we briefly relate the ’t Hooft limit to philosophical discussion of intertheoretic relations. For more details, cf. Bouatta and Butterfield (2012, especially Sects. 1 and 2).

In Butterfield and Bouatta (2011; cf. also Butterfield 2011, especially Sects. 1, 3, 7), we introduced a schema, and a mnemonic notation, for thinking about such relations. We wrote T_b for the ‘better, bottom or basic’ theory, and T_t for the ‘tainted, top or tangible’ theory; (where ‘tangible’ connotes restriction to the observable). Our schema was that in some cases T_t is deduced from T_b (taken together with suitable auxiliary definitions), in some limit of a parameter; and although deduced, T_t exhibits novel features compared with what one sees in T_b—and so deserves the label ‘emergent’. We then argued that phase transitions illustrate this schema: with T_b taken as the statistical mechanics of N constituents; T_t as thermodynamics, taken as describing phase transitions in terms of singularities of thermodynamic quantities; and with the limit being the thermodynamic limit, $N \mapsto \infty$. (For more details, including the distinction (here set aside, to save space) between ‘what happens’ *at* the limit and what happens *on the way to* the limit, cf. Menon and Callender (2013) and Norton (2012).)

The present paper illustrates the same schema. T_t can be taken to be QCD or $\mathcal{N} = 4$ SYM in the ’t Hooft limit. The limit is the ’t Hooft limit. The two main novel features (of course related to each other) are: planarity of diagrams (cf. Sects. 2 and 4.1); and for $\mathcal{N} = 4$ SYM, integrability using spin chains (Sect. 4.2). T_b is of course QCD or $\mathcal{N} = 4$ SYM at finite $N \equiv N_c$ (say, $N = 3!$).

Finally, the significance of this lies largely in the light it sheds on the AdS/CFT correspondence, and the hope it prompts that we might completely understand an interacting quantum field theory.

Acknowledgements It is a pleasure to thank the organizers of the conference, EPSA11, for a most pleasant and stimulating meeting. We also thank a referee; and the editors, not least for their patience. This publication was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation.

References

Beisert, N. (2005). The dilatation operator of N = 4 super Yang-mills theory and integrability. *Physics Reports, 405*, 1–202.

Beisert, N., et al. (2012). Review of AdS/CFT integrability: An overview. *Letters in Mathematical Physics, 99*, 3–32.

Bouatta, N., & Butterfield, J. (2012). On emergence in gauge theories at the ’t Hooft limit. Available online at http://philsci-archive.pitt.edu/9288

Brezin, E., & Wadia, S. R. (Eds.). (1993). *The Large N expansion in quantum field theory and statistical physics: From spin systems to two-dimensional gravity*. Singapore: World Scientific.

Brink, L., Schwarz, J. H., & Scherk, J. (1977). Supersymmetric Yang-mills theories. *Nuclear Physics B, 121*, 77–92.

Butterfield, J. (2011). Less is different: Emergence and reduction reconciled. *Foundations of Physics, 41*, 1065–1135.

Butterfield, J., & Bouatta, N. (2011). Emergence and reduction combined in phase transitions. In J. Kouneiher, C. Barbachoux, & D. Vey. (Eds.), *Proceedings of Frontiers of Fundamental Physics 11 (American Institute of Physics)*. Paris 2010.

Coleman, S. (1985). *Aspects of symmetry*. Cambridge: Cambridge University Press.

Hooft, G. 't. (1974). A planar diagram theory for strong interactions. *Nuclear Physics B, 72*, 461–473.

Maldacena, J. (1997). The large N limit of superconformal field theories and supergravity. *Advances in Theoretical and Mathematical Physics, 2*, 231–252.

Menon, T., & Callender, C. (2013). Turn and face the strange…ch-ch-changes: Philosophical questions raised by phase transitions. In R. Batterman (Ed.), *The Oxford handbook of philosophy of physics*. Oxford: Oxford University Press.

Norton, J. D. (2012). Approximation and idealization: Why the difference matters. *Philosophy of Science, 79*(2), 207–232.

Teper, M. (2009). Large N and confining flux tubes as strings – A view from the lattice. *Acta Physica Polonica B, 40*, 3249–3320.

Part V
Philosophy of the Physical Sciences: Perspectives on Spontaneous Symmetry Breaking

Model Landscapes in the Higgs Sector

Arianna Borrelli and Michael Stöltzner

Abstract This paper provides a still preliminary picture of the model landscape of the Higgs sector both within the Standard Model of Elementary Particle Physics (SM) and beyond (BSM). If one considers it a characteristic feature of models to act as autonomous mediators between theory and data in the sense of Morgan and Morrison, most models in the Higgs sector entertain three types of mediating relationships. First, they mediate between the SM, which has become regarded a well-confirmed theory, and experiment because distilling precise predictions out of the SM requires a specification of parameters. Second, they mediate between BSM physics and the available or presumed experimental signatures by implementing the core ideas behind these often speculative generalizations. Third, Higgs models within BSM physics must reproduce the empirical content of the SM in the low-energy limit to remain consistent with experiment. Due to the speculative nature of the physics BSM, the representative features of the respective Higgs models are complex, and a certain class of models is often kept together by a shared story in the sense of Hartmann.

A. Borrelli (✉)
Interdisciplinary Centre for Science and Technology Studies (IZWT), Bergische Universität Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany
e-mail: ari@drwutzke.de

M. Stöltzner
Department of Philosophy, University of South Carolina, Columbia, SC 29208, USA
e-mail: stoeltzn@sc.edu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_20,

Since March 2010, physicists at the Centre Européene pour la *Recherche Nucléaire* (CERN) in Geneva have been collecting data on the proton-proton collisions occurring inside the Large Hadron Collider (LHC). Among the primary goals of the LHC has been to reach a definitive verdict on the Higgs particle, the long sought-for capstone of the Standard Model of Elementary Particle Physics (SM) and the Higgs mechanism by which the SM particles acquire their masses. And indeed, in July 2012, CERN announced the discovery of a particle signature consistent with the SM Higgs and, simultaneously, excluded further energy ranges for Higgs-like particles to dwell in. The precise properties of the discovered bosonic signature will be investigated in the years to come.

Terminology notwithstanding, the SM is as good a theory as one gets in present-day elementary particle physics where making predictions requires perturbative expansions and renormalization techniques, and where coupling constants and parameters abound, all of whom are eventually determined by experimental data or intertheoretical consistency. An important motivation for this understanding is that the SM's empirical predictions have been confirmed with an impressive precision and that it involves important general principles. Most of these successes will not be influenced by the ultimate composition of the Higgs sector, even though some kind of mass-generation mechanism appears indispensable for the SM to retain its status as a satisfactory theory. Higgs physics aside, most physicists hold that the SM will not be the final word on high-energy physics and the energy range to be investigated by the LHC.

During the two decades since the last experimental discovery in the field, the top quark, theoretical physicists have come up with a large variety of models covering virtually all conceivable outcomes of the LHC experiments. Most of these models assume physics beyond the SM (BSM). Some of them imply that a Higgs particle – or a Higgs-multiplet – must eventually be confirmed by the LHC, others dispense with the need of such a particle altogether. On some accounts, the Higgs particle only appears as a by-product of the mass generation mechanism, while others consider it a primary entity. The intricate nature of the Higgs mechanism is not the only reason why physicists believe that the SM is not the final word. Apart from genuinely theoretical considerations, among them the so-called hierarchy problem, the existence of dark matter, and the integration of the gravitational force strongly suggest a physics BSM. And there is widespread hope that the LHC will also discover traces of such 'new' physics, especially once it reaches higher energies.

The aim of the present paper is to draw a still preliminary picture of the model landscape of the physics in the Higgs sector – where such wording is understood to include models in which there is effectively no Higgs particle or a mass generation mechanism different from the SM – and to derive from it some lessons for current philosophical debates on models in the sciences. Since the Higgs sector has become one of the most booming areas of model-building in present-day physics, there is still too much in flux to aspire at a full-fledged classification from which one could read off in detail the representative features of all models. The goal is rather to distinguish broader classes and specify the relationships between the Higgs models and the levels of theory present in elementary particle physics.

The result will be two-fold and reveal a multi-dimensional model landscape. First, if one considers models as autonomous mediators (MaM) between theory and data – as advocated by Mary Morgan and Margaret Morrison (1999) – most models in the Higgs sector are related not only to the theory SM but also to models or speculative theories BSM. Moreover they partake in the general framework of quantum field theory.[1] These multiple relationships to models and theories situated at different energy scales are important for assessing LHC data because particle detectors are strongly theory-laden: they identify the interesting events through an array of coincidences and throw away almost all events that have or should have been identified by previous experiments (cf. Karaca 2013a).

Second, different models cluster around joint ideas that are not necessarily features of the overarching theories but rather have the character of heuristic principles, most of which transcend the context of a specific model. As does any textbook presentation of the Higgs mechanism, these proposals provide stories – in Stephan Hartmann's (1999) sense – that exert a cohesive function on clusters of models, thus contributing to the cartography of the model landscape. It is important to note that, in various forms, such principles and stories have been present throughout the whole history of the SM and that their instantiation by means of a suitable Lagrangian have shaped the shared techniques of quantum field theory.

The paper is organized as follows: First, we present the Higgs mechanism within the SM and its philosophical critique. Second, we recap the relevant features of present-day debates about models. Third, we show how the SM has developed from a bottom-up model to a widely accepted theory and why this has not in the least discouraged model builders. Fourth, we report on an empirical study of the current model landscape in the Higgs sector. Finally, we discuss the possible consequences of the models' mediating role and their complex representative features on the ontological considerations concerning the Higgs mechanism.

1 Models, Mechanisms, and Theories in the Higgs Sector

The SM is based on a product of three gauge symmetry groups $\mathrm{SU}(3) \times \mathrm{SU}(2)_{\mathrm{L}} \times \mathrm{U}(1)$ instantiating the strong and the electroweak forces. The mathematical techniques involved are encoded in a set of physically coherent and, at least partly, mathematically rigorous rules which have been developed over decades, most prominently the renormalization group. The SM has been numerically confirmed to an astonishing extent, and the search for the remaining quarks has resembled the

[1] This framework may be given by a fully axiomatized theory – which facilitates the work of mathematical physicists and philosophers alike – or by a theory in the mathematically looser sense used among theoretical physicists. Even though one of us has his views (cf. Stöltzner 2012), the present paper can remain uncommitted about the hotly debated question as to which approach is 'more natural' in a philosophical perspective (cf. Fraser 2011; Wallace 2011).

filling up of empty slots in Mendeleev's table – a philosopher's textbook case of successful explanation and a case in point for scientific realism. This remarkable success of the SM, including the discovery of all particles mediating the strong and electroweak interactions, has prompted philosophers of science to address the ontological aspects of gauge symmetries as the basic theoretical entities of our world. But there is still a major lacuna.

All this intricate theoretical architecture requires some non-trivial method for introducing masses. This is best explained in a historical perspective.[2] The unification of electromagnetic and weak interactions by means of a local gauge invariant theory had first been proposed in the early 1960s. However, the phenomenology of weak interactions required the introduction of massive vector bosons, and their mass terms broke the local gauge invariance, spoiling the renormalizability of the theory and leaving physicists with a theory devoid of predictive power. The solution was believed to lie in a notion borrowed from superconductivity, to wit, spontaneous symmetry breaking (SSB). However, the Goldstone theorem showed that SSB yielded massless, spinless particles ("Goldstone bosons"), the likes of which had never been observed. It took the parallel efforts of many scientists to realize in the following years that, if the spontaneously broken symmetry is a local gauge invariance, the Goldstone bosons can be formally eliminated by a redefinition of the fields before (second) quantization. In this process, the Lagrangian is transformed into a new one which contains no Goldstone bosons, but instead mass terms for the gauge vector bosons. What is more, such a Lagrangian is renormalizable. The names associated with this discovery are plenty (Schwinger, Andersons, Englert, Brout, Kibble, Guralnik, Hagen, Higgs), but it has come to be referred to as the "Higgs mechanism".

Over the years, the Higgs mechanism has repeatedly come under criticism from physicists and philosophers alike. To begin with, it appears as an ad hoc strategy devised to solve one and only one problem within the SM by introducing an additional entity. Yet according to most theoretical models the Higgs particle is identified by its decay chain, as have been all other particles of the SM. There are quite a few possible channels, some of which are more accessible for the LHC than others. Honest worries about ontological parsimony notwithstanding, the Higgs particle is thus not causally isolated. Still it is the product of a complex mechanism based on concepts of distinct origins.

Philosophers of science are mainly worried about the ontological status of the Higgs particle and the explanatory depth of the Higgs mechanism. Holger Lyre, for one, has argued that the above-sketched transformation of the Lagrangian "consists in a mere re-shuffling of degrees of freedom [... which does not suggest a] straightforward ontic interpretation of the Higgs mechanism." (2008, p. 119) Yet it enjoys "the same ontological status like any other mechanism of spontaneous symmetry breaking [...] in ferromagnets or superconductors." (Wüthrich 2012, p. 6) But, as Lyre (2012) retorts, the SM provides no causal story that links the

[2] A more detailed study of the emergence of the Higgs mechanism is Karaca (2013b). For a historical overview, see Pickering (1984), Hoddeson et al. (1997), and t'Hooft (2005).

unbroken and the broken phases. In the same vein, Morrison has diagnosed that the textbook narrative "lacks the kind of independent evidence capable of supporting a realist interpretation of SSB." (2003, p. 361)

To John Earman's mind: "Philosophers of science should be asking [. . .]: What is the objective (i.e. gauge invariant) structure of the world corresponding to the gauge theory presented in the Higgs mechanism?" (2004, p. 1239) He recommends a rigorous mathematical approach starting from algebraic quantum field theory and to understand the gauge conditions as imposing constraints. In this way one might hope to factor out the unphysical degrees of freedom and arrive at a gauge-invariant structure. But the prospects of Earman's program are mixed. Although Ward Struyve (2011) has recently implemented Earman's program by providing a manifestly gauge invariant formulation of the Higgs mechanism at the classical level, its counterpart at the quantum level will most likely not arise as a straightforward extension of this formulation. For results obtained within the algebraic approach point to significant differences between the classical and the quantum level, in virtue of which the non-regular state implementing the gauge constraint does not arise from an operator that is simply the quantization – in the sense of the correspondence principle – of a classical Higgs field. (cf. Stöltzner 2012) The Higgs mechanism accordingly remains pretty elusive, as Chris Smeenk (2006) has aptly put it. Accordingly, the present paper, in the same vein as Smeenk's, takes a bottom-up approach combining empirical and analytical tools. It suggests that the representative features of models may serve a first step to ease ontological worries concerning the Higgs sector.

2 Models as Autonomous Mediators and their Stories

For the philosophical analysis, we are relying upon the approach of Morgan and Morrison that understands models as mediators between theory and empirical data (MaM) and endows them with autonomous representative features. This "autonomy is the result of two components (1) the fact that models *function* in a way that is partially independent of theory and (2) in many cases they are *constructed* with a minimal reliance on high level theory." (1999, p. 43) Sometimes, they even function as measuring devices for specific parameters of the problem at hand that are typically not given by the framework theory. Adopting the MaM approach does not imply that we consider a semantic account in the style of Giere (1999) inapplicable to particle physics. After all, Morgan and Morrison primarily criticize Giere as too closely focused on theory – which should be less of an issue in present-day particle physics. The problem is rather that Giere, on a psychological basis, introduces a two-dimensional map of models in which the vertical levels are globally separated. However, such a global gradation does not exist for the physics BSM, which involves a variety of different energy scales that are not necessarily separated – as are the physics of the atom, the physics of the nucleus, and the sub-nuclear world of high-energy physics.

Morrison's prototype example is the boundary layer model of a flow in a pipe. (Ironically, at the height of its career in the early twentieth century, it was, more often than not, called a theory.) It divides the flow into an ideal flow in the center of the pipe and a two-dimensional, friction-dominated boundary layer. Both regimes can be understood as simplifications of the intractably complex Navier-Stokes equations that govern the flow as a whole. However these simplifications do not arise in a natural or canonical way from the mathematics of the Navier-Stokes theory, but require the choice of a model that involves genuinely representative features.

Representation, it is true, comes at a price because it restricts a model to a certain class of phenomena. Compare the initial phenomenological models of the atomic nuclei. While the liquid drop model is able to explain most instances of nuclear fission – in analogy to the splitting of a droplet – by a phenomenological formula for the energy, the shell model adapts the quantum mechanics of the atomic hull for a model of the nucleus. The shell model can explain why nuclei consisting of certain numbers of neutrons and protons that correspond to closed shells, are more stable than others. (cf. Morgan and Morrison 1999, pp. 23–24) Even if one starts, at a theoretically deeper level, from quantum chromodynamics (QCD), one can build more than one autonomous phenomenological model of the nucleus emphasizing distinct theoretical properties, among them (spatial) confinement of the quarks or chiral symmetry (and its dynamical breaking). Using this case study, Hartmann (1999) interprets models as stories in which the representative features emerge as consequences of the respective theoretical agendas.

> A story is a narrative told *around* the formalism of the model. It is neither a deductive consequence of the model nor of the underlying theory. It is, however, *inspired* by the underlying theory (if there is one). This is because the story takes advantage of the vocabulary of the theory (such as 'gluon') and refers to some of its features. [...] Using more general terms, the story fits the model in a larger framework (a world picture) in a non-deductive way [...]; it complements the formalism. (1999, p. 344)

An important point is that stories permit one to operate with models in a highly theoretical domain, such as present-day particle physics. Morgan (2001) has argued that even broader narratives are indispensable in order to apply models to real-world phenomena, thus extending Hartmann's stories. However this additional narrative dimension lies outside the scope of the present paper because it involves the relationship between models of the object processes and models – and computer simulations – of the detector. These are, it is true, indispensable to produce any experimental data in high-energy physics and to actually test theoretical models, but they are not part of the Higgs model landscape shared by theorists and experimentalists.

While the Navier-Stokes equations still encompass the whole domain of hydrodynamics, including the boundary layer model, the models of present-day particle physics are not situated beneath a sufficiently specific overarching theory – whatever its predictive power might be. Of course, particle physicists operate within the theoretical framework of quantum field theory, adopt its general principles and calculation strategies, which permits them to build models by means of a Lagrangian instantiating the basic gauge symmetry and treat it by a set of well-entrenched techniques. While proponents of the algebraic approach believe that the framework of quantum field theory can be sufficiently specified to reach at least general conclusions about the SM, others consider such a goal unrealistic. This does not mean that advocates of conventional quantum field theory – to adopt Wallace's (2011) terminology – necessarily take an exclusively instrumentalist attitude towards their research program. Too many basic concepts of quantum field theory are as well defined as the theoretical physicist can wish, some even have a solid mathematical foundation. Moreover, the framework of quantum field theory can be applied to larger domains of physics, among them condensed matter physics. The point is rather that at present no version of quantum field theory is in such a shape that someone would argue that it is only complexity that prevents a deduction of the physical phenomena, which at least in principle would be possible.

The SM itself has arisen bottom-up from a specific Lagrangian. Some physics BSM is built up simply by modifying this Lagrangian while other models involve new basic concepts, such as a new fundamental symmetry. All of them take the SM as a reference theory, however not as a sufficiently specific and universally accepted theory at the higher level but as a consistency constraint for model building. For all models must reproduce its predictive content in the low-energy limit. As a matter of fact, it took considerable time until the SM earned itself the rank of a theory.

3 From a Model to a Theory and Back to Models Again

The history of the SM began with two papers by Steven Weinberg and Abdus Salam in the early 1960s, each independently proposing a model of electroweak interactions expressed by the gauge group $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)$. The local gauge invariance was broken spontaneously by a scalar field through the Higgs mechanism. Their hope was that SSB would ensure renormalizability. Yet it took a while until Gerard t'Hooft showed how to renormalize the Weinberg-Salam model, prompting an increased interest into quantum field theory among mainstream theoretical physicists.

The phenomenological successes of the Weinberg-Salam model and QCD, hence both parts of the SM, have been considerable: the discovery of neutral currents (1973), the explanation of Bjorken scaling in deep inelastic lepton-hadron scattering (1973), the prediction and discovery of charm quarks (1974), the indirect (1983) and direct (1989) detection of electroweak vector bosons $\mathrm{W}^{\pm}$ and Z at the LEP experiment, the predecessor of LHC.

Yet this abundance of theoretical and experimental support, the absence of any anomalies or unexplained experimental signatures, and its increased recognition as a theory did not make the SM immune against attempts to alter, extend or reject various parts of it. Model-building BSM, e.g. by modifying its Higgs sector, was hardly ever prompted by the necessity of matching the SM better to the phenomena, but rather by the desire to explore its theoretical and phenomenological potentialities, e.g. by adding some new symmetry or reduce the number of free parameters. For, initially, the success of the SM and its Higgs mechanism were regarded more as successes of the various theoretical ingredients contained in it (local gauge symmetry, renormalizability, spontaneous symmetry breaking), than as a confirmation of the predictions of that specific combination. It took considerable time until the SM was elevated to its peculiar rank as a theory, which ultimately yielded a clear distinction between SM and BSM physics, the latter exploring allegedly more fundamental theories dwelling on higher orders of energy.[3]

4 Model Landscapes in the Higgs Sector

In a recently launched research project 'Epistemology of the LHC', one of us (A.B.) has undertaken a real-time mapping of the model dynamics based on the preprints on the internet platform arxiv.org. Practically all papers in contemporary elementary particle physics are deposited there before they are submitted to journals or conference proceedings. Since February 2010, all preprints appearing in arxiv:hep-ph (devoted to high-energy phenomenology) or cross-posted there from 'hep-th' (containing more speculative theories) are examined and those are selected that discuss the Higgs sector of the SM or modify it involving BSM physics in order to explain mass generation. The preprints thus individuated are analyzed in more detail and tentatively subdivided into a number of categories according to the models discussed in them.

For the period from February, 2nd, 2010 until October 7th, 2010, it was found that, among 422 relevant preprints, 38 were treating the Higgs sector within the SM. Among those involving BSM physics, about 90 % could be subsumed under five major categories. Most of the remaining papers can be accommodated within a couple of minor classes, but a handful could not be assigned to any class. A second screening in 2011 only found small changes in the relative percentages, neither the dissolution of one of the five categories nor the emergence of an entirely new category. The influence of the recent discoveries of LHC will be the object of a third screening period.

1. *Models with additional Higgs-fields* (75 papers): The SM contains the minimum number of Higgs fields necessary to realize the Higgs mechanism. The presence of additional Higgs fields does not change the fact that SSB is implemented through the Higgs mechanism, but the increased number of free parameters makes it easier to obtain values for the Higgs masses that fit experimental limits. In addition to the neutral SM Higgs particle, charged Higgs bosons appear.
2. *Supersymmetry* (108 papers): Supersymmetry transforms bosons into fermions and vice versa. Since, however, no such symmetry has been observed even

[3] For an overview of the current state of BSM physics, see Bustamante et al. (2010), Rattazzi (2006) and Altarelli (2010).

in approximate form, supersymmetry is assumed to be broken so that the supersymmetric partners of all known particles have such high masses that they could not have been observed by the existing experiments. While the LHC has meanwhile excluded further regions of the energy spectrum, the observed mass of the Higgs candidate is still consistent with supersymmetry. One major characteristic differentiating the various supersymmetric models is the way in which supersymmetry is broken. Due to the requirement of (approximate) supersymmetry, more Higgs fields have to be introduced to break electroweak symmetry than in the SM. As in the case of (1), the additional fields translate into charged Higgs particles.

3. *Models with dynamical SSB* (74 papers): Other than categories (1) and (2), these models do not contain an elementary Higgs field and accordingly no Higgs mechanism. The core idea is to introduce additional elementary fields, usually fermions with a behavior similar to that of quarks. The new particles combine into bound states, which spontaneously break the symmetry. This kind of SSB is called dynamical because it is due to the dynamical interactions of elementary fields. Some models assume that one of the symmetry-breaking bound states is the Higgs boson (technicolor), but others dispense with any Higgs-like particle at all ('Higgs-less' models).[4]
4. *Higgs as pseudo-Goldstone boson* (23 papers): Although the Higgs boson is an elementary particle breaking local gauge symmetry through the Higgs mechanism, just as in the SM, the Higgs boson itself is regarded as the product of the spontaneous breaking of a symmetry, a fact expressed technically by saying the Higgs is a 'pseudo-Goldstone boson'. The details, among them which symmetry is associated to this boson, vary greatly.
5. *Extra dimensions* (61 papers): An approach which to an increasing extent has attracted attention since 2000 is based on the intuition that space-time has more than four dimensions. Both the idea and the formal tools were imported from string theory. This category is by far the most colorful one, and it can be further subdivided, for instance according to the number of extra-dimensions or how they are made invisible. SSB is realized in different ways: some models contain four-dimensional Higgs fields and use the Higgs mechanism, while in others the Higgs field is the 'shadow' of a field moving in extradimensional space. It is also possible to spontaneously break local gauge symmetry without a Higgs field by using the way in which the extra-dimensions are 'rolled up' to make them nonobservable. Such 'Higgs-less' models are at times regarded as equivalent to category 3.

The classification just provided, despite being stripped of most details, still cannot claim to be mutually exclusive and exhaustive. For example, some models with dynamical symmetry breaking still introduce an elementary Higgs field and

[4] For a historical study of these Higgs models, see Borrelli (2012).

the Higgs mechanism, so that spontaneous breaking of electroweak symmetry takes place at more than one stage. Moreover in the last decade, there has been a growing tendency to combine different approaches into a single model, for example supersymmetry and extra dimensions, or dynamical symmetry breaking and the Higgs as pseudo-Goldstone boson. This underscores that the classification does not group together models deduced from some common principle or theory, but rather models which represent in different ways the same desired features, such as supersymmetric partners or a SSB mechanism. These stories combine easier because, unlike formal theories, stories do not succumb to strict mathematical consistency requirements – as long as they are consistent with the larger theoretical framework provided by quantum field theory, which in effect allows for the formulation of a large variety of effective field theories instantiating a story. Thus these categories have to be conceived as clusters of models which, while sharing one or more overarching properties, at the same time also embed very different elements of diverse provenance and try to integrate different stories into a coherent overall narrative.

A feature common to all Higgs models is the fact that, in the limit of low energies, they recover the experimentally confirmed empirical content of the SM. Because of this, all models of BSM physics and the related stories are connected to the SM, but they are not derived or theoretically dependent upon it. For they introduce new elements which are supposed to manifest themselves only at energies above the SM. These extensions are thus underdetermined by the standard model and the currently available empirical evidence, even though their space is constantly reduced by the LHC data. This underscores that the SM as the leading and empirically best corroborated theory in the field acts primarily as consistency constraint and a conceptual inventory for model builders.

5 Conclusion

Let us summarize in what sense the models in the Higgs sector exhibit the autonomy demanded by the MaM approach both as regards the history of construction and establishment of the SM (Sect. 3) and as regards their present function (Sect. 4). This autonomy typically comes with a story because, unlike Prandtl's water tunnel, the LHC does not simultaneously represent a model *and* an experimental device. Rather do LHC physicists consider the model landscape of Higgs physics and the detector models as two separate dimensions.

The historically developed construction, or the standard textbook version of the Higgs mechanism, is not the only story around. Neither the alternative methods to realize electroweak symmetry breaking, nor the various ideas forming the core of BSM approaches are derived from the SM or some general quantum field theory by making approximations or simplifications. The Higgs – or Higgs-less – models are constructed in autonomous processes mathematically implementing certain heuristic ideas or general principles, among them supersymmetry or dynamical symmetry breaking, by using standard techniques of quantum field theory. These ideas come with stories motivating or even justifying the respective model. Some of these principles and their associated stories have entered high energy physics from neighboring domains, both well-established ones like solid state physics and speculative ones like string theory.

As regards their functions, Higgs models, first, reproduce the phenomenology of the SM, thus showing how that theory can be extended or modified without losing its predictive power. Second, the various models have long managed to live with the narrowing bounds on the mass of the Higgs boson posed by the increasing experimental data by changing the predictions for the masses of the particles populating the Higgs sector, so as to provide a higher probability for them to fall within the limits allowed by past experiments. Third, Higgs models often eliminate the need for fine-tuning in the renormalization of the Higgs mass.

As regards their representative features, the Higgs models refer to different energy scales and stand in three kinds of mediating relationships. First, they mediate between the SM and the experimental data pretty much in the style as described by MaM, especially in cases where the SM cannot produce exact predictions because of the uncertainty in the values of its basic parameters. Second, they mediate between BSM physics and the data by precisely implementing additional theoretical ingredients of various kinds, by making explicit the stories that motivate the respective model. Third, the fact that Higgs models within BSM physics reproduce the SM predictions in the low-energy limit functions as a consistency constraint that restricts model-builders' autonomy.

Due to the second type of mediating relationship, the representative features of some Higgs models BSM may be opaque and strongly rely on their motivating story. But these stories are not fictional. For despite the sometimes indirect motivations for and the speculative character of certain parts of BSM physics, model-builders usually regard them as candidates for fundamental features of reality. Thus, BSM models are assigned representative features not only by their reproduction of confirmed SM phenomenology but also by their capability to positively address a number of general issues which are believed to be indicative of a more fundamental level of physical reality dwelling at higher energies. Physicists, it appears, prefer the stories to be real and ultimately develop into textbook narratives motivating a well-confirmed theory.

Rather than accepting such ontological commitments at face value, philosophers should consider them as an expression of the representative features of the models concerned. These representative features, as complex as they are, may provide a better bottom-up therapy to address philosophers' well-justified ontological worries about the Higgs mechanism than a head-on approach. For physicists have both ontological commitments and an exploratory attitude that, in the theoretical realm, is well captured by the notion of a story.

Acknowledgement This work is part of the Wuppertal-based research group "Epistemology of the Large Hadron Collider" supported by the Deutsche Forschungsgemeinschaft (DFG); http://www.lhc-epistemologie.uni-wuppertal.de/.

References

Altarelli, G. (2010). Particle physic at the LHC start. arXiv:hep-ph/1010.5637v1.

Borrelli, A. (2012). The case of the composite Higgs: The model as a "Rosetta stone" in contemporary high-energy physics. *Studies in the History and Philosophy of Modern Physics, 43*, 195–214.

Bustamante, M., Cieri, L., & Ellis, J. (2010). Beyond the standard model for montañeros. arXiv:hep-ph/0911.4409v2.

Earman, J. (2004). Laws, symmetry, and symmetry breaking: Invariance, conservation principles, and objectivity. *Philosophy of Science, 71*, 1227–1241.

Fraser, D. (2011). How to take particle physics seriously: A further defense of axiomatic quantum field theory. *Studies in History and Philosophy of Modern Physics, 42*, 126–135.

Giere, R. (1999). *Science without laws*. Chicago: University of Chicago Press.

Hartmann, S. (1999). Models and stories in hadron physics. In M. Morgan & M. Morrison (Eds.), *Models as mediators: Perspectives on natural and social science* (pp. 326–346). Cambridge: Cambridge University Press.

Hoddeson, L., Brown, L., Riordan, M., & Dresden, M. (Eds.). (1997). *The rise of the standard model: Particle physics in the 1960s and 1970s*. Cambridge: Cambridge University Press.

Karaca, K. (2013a). The strong and weak senses of theory-ladenness of experimentation: Theory-driven versus exploratory experiments in the history of high-energy particle physics. *Science in Context, 26*, 93–136.

Karaca, K. (2013b). The construction of the Higgs mechanism and the emergence of the electroweak theory. *Studies in History and Philosophy of Modern Physics, 44*, 1–16.

Lyre, H. (2008). Does the Higgs mechanism exist? *International Studies in the Philosophy of Science, 22*, 119–133.

Lyre, H. (2012). The just-so Higgs story: A response to Adrian Wüthrich. *Journal for General Philosophy of Science, 43*, 289–294.

Morgan, M. (2001). Models, stories, and the economic world. *Journal of Economic Methodology, 8*, 361–384.

Morgan, M., & Morrison, M. (Eds.). (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge: Cambridge University Press.

Morrison, M. (2003). Spontaneous symmetry breaking: Theoretical arguments and philosophical problems. In K. Brading & E. Castellani (Eds.), *Symmetries in physics: Philosophical reflections* (pp. 347–363). Cambridge: Cambridge University Press.

Pickering, A. (1984). *Constructing quarks – A sociological history of particle physics*. Chicago: The University of Chicago Press.

Rattazzi, R. (2006). Physics beyond the standard model. arXiv:hep-ph/0607058v1.

Smeenk, C. (2006). The elusive Higgs mechanism. *Philosophy of Science, 73*, 487–499.

Stöltzner, M. (2012). Constraining the Higgs mechanism: Ontological worries and the prospects for an algebraic cure. *Philosophy of Science, 79*, 930–941.

Struyve, W. (2011). Gauge invariant accounts of the Higgs mechanism. *Studies in History and Philosophy of Modern Physics, 42*, 226–236.

't Hooft, G. (Ed.). (2005). *50 years of Yang-Mills theory*. Singapore: World Scientific.

Wallace, D. (2011). Taking particle physics seriously: A critique of the algebraic approach to quantum field theory. *Studies in History and Philosophy of Modern Physics, 42*, 116–125.

Wüthrich, A. (2012). Eating Goldstone bosons in a phase transition: A critical review of Lyre's analysis of the Higgs mechanism. *Journal for General Philosophy of Science, 43*(2), 281–287.

Practical Unification of Solid-State and Particle Physics in the Construction of the Higgs Mechanism

Koray Karaca

Abstract A number of philosophical accounts have been offered as to what sort of unification exists between the electromagnetic and weak interactions in the Glashow-Weinberg-Salam electroweak theory of elementary particle physics. In this paper, unlike the previous studies, I seek to address how "unity" in science might be interpreted in view of the construction process of the Higgs mechanism, which was a decisive step in the construction of the electroweak theory.

1 Introduction

In philosophical literature, scientific unity has been typically understood as "theoretical unification" and considered to take place by virtue of various formal relations involved in the structures of theories (see Cat 2007). Oppenheim and Putnam's "unity as reduction" thesis (1958) has been a prominent account of theoretical unification in the second half of the past century. In this account, unity is taken to consist essentially in *inter-theoretic* reductions between the different levels of a hierarchy taken to represent the order of the part-whole relations among the entities in the world. Over the last few decades, however, non-reductive accounts of unity have also been advanced. Contrary to the unity as reduction thesis, those accounts have recognized that scientific unity is possible without inter-theoretic reductions. A prominent account of non-reductive unification has been offered by Darden and Maull (1977), who have shown that unity as reduction thesis fails to account for some important cases in biology. Darden and Maull do not argue against the possibility that inter-theoretic reductions might occur in some possible

K. Karaca (✉)
Interdisciplinary Centre for Science and Technology Studies (IZWT), University of Wuppertal, Gauss str. 20, 42119 Wuppertal, Germany
e-mail: karacak@gmail.com

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_21,

cases. Rather, they propose that unity can come in degrees and occur when two different fields are linked to each other through what they call an "interfield" theory; e.g., the chromosome theory of Mendelian heredity bridging the fields of cytology and genetics, and the operon theory relating the fields of genetics and biochemistry. In Darden and Maull's account, interfield theories set out and explain various theoretical connections – e.g., conceptual, ontological, and explanatory connections – that exist between fields (p. 48). Even though Oppenheim and Putnam's and Darden and Maull's accounts disagree on the way unity takes place in scientific practice, both accounts conceive of unity as theoretical unification; in that while the former takes unity to consist in inter-theoretic reductions, the latter takes it to consist in theoretical interconnections between different fields of a branch of science.

Recently, the prevalence of theoretical unification has been challenged by Cat (1998, 2006), who has pointed out that unity can *also* lie outside the mathematical structures of scientific theories and consist in borrowing, sharing, and circulation of physical concepts, methods, tools, and techniques among different sub-disciplines. In a similar vein, Grantham (2004) has drawn attention to what he calls "non-theoretical (practical) interconnections" between different fields of a branch of science. Grantham explicates those practical interconnections as follows (p. 143):

Heuristic dependence. The theories and/or methods of one field can guide the generation of new hypotheses in a neighboring field.

Confirmational dependence. The methods and/or data of one field may be used to confirm hypotheses generated in a neighboring field.

Methodological integration. Methods can be developed to assess an hypothesis in light of the data (often generated by distinct methods) of two fields.

Grantham acknowledges that Darden and Maull's model adequately accounts for some important cases in scientific practice. But, he suggests that unity between fields can also consist in *practical interconnections* that, unlike theoretical interconnections, do not require interfield theories. Grantham calls this type of unity "practical unification" and points out that, unlike reductive unity, it comes in degrees and is largely *independent* of theoretical unification (pp. 144–145), which primarily concerns the structures of theories and associated formal theoretical relations. Grantham does not argue against theoretical unification, either in the form of inter-theoretic reductions or interfield theories; rather his suggestion is that practical and theoretical unifications should be understood as two different, but not necessarily mutually exclusive, ways in which unity may occur in scientific practice.

Grantham has illustrated how the practical unification of fields can take place in the form of the methodological integration of two fields with a case study concerning the fields of paleontology and neontology, which respectively study fossils and living organisms. In this paper, I shall illustrate how the practical unification can take place in the form of the heuristic dependence of one field on the other in the case of the construction process of the Higgs mechanism, which I shall describe in the next section.

2 A Short Guide to the Construction of the Higgs Mechanism

The concept of "spontaneous symmetry breaking" is used in physics to characterize the situation in which the "laws of a theory may describe the behavior of a system that is not itself symmetric under a transformation that is a symmetry of those laws" (Healey 2007, p. 170). This concept was first introduced into the context of elementary particle physics by Nambu's quantum field theoretic elucidation of the Bardeen-Cooper-Schrieffer (BCS) theory of superconductivity (Bardeen et al. 1957) in solid-state physics (Brown and Cao 1991), where the derivation of the *Meissner effect* was not originally gauge invariant. The Meissner effect is the phenomenon that magnetic field is expelled from a superconductor's surface and can only penetrate a very small length. According to the BCS theory, the Meissner effect results from the formation of "Cooper pairs" below a critical temperature in a superconductor. Cooper pairs are electron pairs that have equal and opposite spins and momenta. In the language of solid-state physics, the interaction between the Cooper pair electrons is mediated through "phonons", which are conceived to be massless and spinless (i.e., bosonic) energy excitations resulting from the vibrations of ions in a lattice. The wave-functions of the Cooper pair electrons have *long range phase coherence*, resulting in the spontaneous breaking of the *global* U(1) phase symmetry exhibited by a system of electrons in a lattice before the formation of Cooper pairs. Here, global phase symmetry is the invariance of the electron wave-function under the following transformation: $\psi \rightarrow \psi e^{i\theta}$, where θ is an arbitrary real constant. It is also to be noted that, in quantum field theory, *global* gauge symmetry is the invariance of the Lagrangian under a continuous group of transformations that are the *same* at every point in space and time; whereas *local* gauge symmetry is the invariance of the Lagrangian under a continuous group of transformations that are space and time dependent.

Nambu (1960) demonstrated that, in the case of the Meissner effect, by virtue of the existence of phonon states, there exist *generalized forms* of the "Ward-Takahashi identity": $k_\mu M^\mu(k) = 0$, where the four-vectors represent respectively the momentum of the photon involved in a scattering process and the associated scattering amplitudes.[1] In quantum field theory, the Ward-Takahashi identity is a statement of "current conservation" as a consequence of local gauge invariance (see, e.g., Peskin and Schroeder 1995, Sect. 7.4). Therefore, Nambu's result (1960) indicated that the local U(1) gauge invariance of electromagnetism is preserved in a superconductor by virtue of the existence of phonon states. Here, the local U(1) gauge invariance of electromagnetism is the invariance of Maxwell's equations under the local U(1) gauge transformation of the electromagnetic four-vector potential: $A_\mu(x) \rightarrow A'_\mu(x) = A_\mu(x) - \partial_\mu \Lambda(x)$, where $\Lambda(x)$ is an arbitrary scalar function of space and time.

[1] Throughout the paper, all indices run from 0 to 3, and the metric signature is taken as (–1, +1, +1, +1).

A year later, Nambu (1961), together with Jona-Lasinio, constructed a *composite* model of nucleons based on an analogy drawn to the BCS theory. In the same way as the energy-gap in a superconductor is brought about by the interaction between fermion pairs, i.e., the Cooper pair electrons, in the Nambu – Jona-Lasinio model the observed mass of the nucleon is taken to be a *collective* effect that is brought about by some *unknown* interaction between *massless* fermion pairs. As electrons in a superconductor pair up to form Cooper pairs in the BCS theory, in the Nambu – Jona-Lasinio model, massless fermions pair up to form fermion pairs as a result of an *unknown* interaction. This in turn results in the spontaneous breaking of the *global* U(1) chiral symmetry, which is the invariance of the interaction between massless fermions under the global chiral transformation: $\Psi \rightarrow \Psi e^{\mathrm{i}\alpha\gamma_5}$ where $\mathit{\Psi}$, γ_5 and α denote respectively the fermion field, the chirality operator (i.e., the fifth Dirac matrix), and an arbitrary real constant. As the spontaneous breaking of the global U(1) phase symmetry brings about phonons as massless spinless collective excitations in a superconductor in the BCS theory, the spontaneous breaking of the global U(1) chiral symmetry brings about massless pseudoscalar bosons in the form of massless pions in the Nambu – Jona-Lasinio model.

The Nambu – Jona-Lasinio model suggested the possibility that the "mass problem" of gauge theories of the Yang-Mills type might be solved through spontaneous symmetry breaking. The mass problem stems from the technical fact that mass terms in quantum field theory are quadratic in fields (not containing derivatives) and thus are *not* gauge invariant. This means that the Lagrangian of any gauge invariant quantum field theory of the Yang-Mills type should contain *no* mass terms quadratic in gauge fields (not containing derivatives). However, all experimental evidence indicates that the only *massless* vector boson is the photon that mediates the electromagnetic interaction, suggesting that all other vector bosons in gauge theories of the Yang-Mills type should be *massive*. Therefore, it remains to be answered how vector bosons can acquire mass in these theories without destroying gauge invariance.

In 1961, Goldstone (1961) conjectured that the spontaneous breaking of a continuous *global* symmetry in a Lorentz-covariant Lagrangian brought about massless spinless bosons – often referred to as the "Goldstone bosons" – like the massless pions in the Nambu – Jona-Lasinio model. Goldstone's conjecture was subsequently proved in a joint paper by Goldstone, Salam, and Weinberg (1962), and it was elevated to the status of a theorem, which has come to be referred to as the "Goldstone theorem." Since there was no experimental evidence whatsoever about the existence of massless spinless bosons, the Goldstone theorem cast doubts on the possibility that the mass problem might be solved through spontaneous symmetry breaking.

In a paper published in 1963, the solid-state physicist Anderson (1963) questioned the applicability of the Goldstone theorem in the context of solid-state physics, in particular in the case of superconductivity. Anderson drew upon an argument previously put forward by Schwinger, that "the gauge invariance of a vector field does not necessarily imply zero mass for an associated particle if the current vector coupling is sufficiently strong" (1962, p. 397). Schwinger's argument was based on the observation that *local* gauge invariance does not preclude gauge quanta to be massive, if the vacuum polarization tensor has a pole (i.e., singularity) at momenta $p^2 = 0$ (see Peskin and Schroeder 1995, pp. 245–246). Schwinger demonstrated that the existence of such a pole is possible under the condition of conserved current-vector field coupling. He illustrated this argument in a two-dimensional (one time and one space dimension) model of quantum-electrodynamics, where, due to conserved current-vector field coupling, the polarization tensor develops a pole at $p^2 = 0$; thereby the photon acquires mass.

Anderson (1963) relied on the plasmon theory of the free-electron gas (see Nozieres and Pines 1958) and suggested the following explanation for the Meissner effect. Inside a superconductor, longitudinally polarized massless phonons, which result from the spontaneous breaking of the global U(1) phase symmetry and which mediate between the Cooper pair electrons, combine with transversely polarized massless photons which mediate the magnetic interaction. As a result of this coupling, massless photons acquire longitudinal polarization state and thus become *effectively* massive,[2] thereby causing the magnetic interaction to become a short-range interaction inside a superconductor. Note that, in quantum field theory, a massless vector boson has only two transverse polarization states; whereas a massive vector boson has, in addition, one longitudinal polarization state due to its mass. Therefore, Anderson's above explanation of the Meissner effect indicates that as a result of the coupling between the external magnetic field and the Cooper pair current, which is conserved in a superconductor, massless photons acquire longitudinal polarization state and thus become massive. Anderson thus pointed out that the Meissner effect illustrated Schwinger's suggestion that as a consequence of conserved current-vector field coupling vector bosons might not be necessarily massless.

In the same paper, Anderson (1963) drew an analogy to the above-stated explanation of the Meissner effect. Regarding massless phonons as the Goldstone bosons in a superconductor, Anderson suggested that in the same way as massless phonons and massless photons combine and form massive plasmons as a result of conserved current-vector field coupling in a superconductor, in gauge theories of the Yang-Mills type the Goldstone field and the massless vector field would combine through a conserved current-vector field coupling to form a massive vector field. In Anderson's own words:

> It is likely, then, considering the superconducting analog, that the way is now open for a degenerate-vacuum theory of the Nambu type without any difficulties involving either zero-mass Yang-Mills gauge bosons or zero-mass Goldstone bosons. These two types of bosons seem capable of "canceling each other out" and leaving finite mass bosons only … The only mechanism [I suggest] for giving the gauge field mass is the degenerate vacuum type of theory, in which the original symmetry is not manifest in the observable domain. (1963, p. 441)

[2] These "massive" photons are called in solid-state physics "plasmons," which are regarded as *collective excitations* of the free-electron gas in a metal.

Anderson's above analogical argument is grounded on the consideration that gauge theories of the Yang-Mills type have *local* gauge invariance. In these theories, the condition of conserved current-vector field coupling – in the way suggested by Schwinger (1962) – is ensured by the procedure of "minimal coupling", which serves to transform a global gauge invariant *free-field* Lagrangian into a local gauge invariant Lagrangian (for technical details, see, e.g., Peskin and Schroeder 1995, Chap. 4). Therefore, Anderson's argument can be interpreted as suggesting that, in a local gauge invariant theory, it is possible both to give mass to vector bosons and to get rid of the unwanted Goldstone bosons through spontaneous breaking of local gauge symmetry, as opposed to in a theory with *only* global gauge invariance, where, due to the absence of conserved current-vector field coupling, the Goldstone bosons find no chance to transform away from the particle spectrum of the theory.

Anderson's proposal was taken up by Higgs (1964), who examined the spontaneous breaking of the local U(1) gauge symmetry in a simple classical (*unquantized*) theory where two real scalar fields φ_1 and φ_2 interact with a real vector field A_μ in the way represented by the following Lagrangian:

$$\mathcal{L} = -\frac{1}{2}\left(\nabla_\mu\varphi_1\right)^2 - \frac{1}{2}\left(\nabla_\mu\varphi_2\right)^2 - V\left(\varphi_1^2 + \varphi_2^2\right) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{1}$$

where $\nabla_\mu\varphi_1 = \partial_\mu\varphi_1 - eA_\mu\varphi_2$; $\nabla_\mu\varphi_2 = \partial_\mu\varphi_2 + eA_\mu\varphi_1$; $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$; V denotes the potential energy of the scalar fields; and e is a dimensionless coupling constant. It is important to note that, in conformity with Anderson's proposal (1963), $\mathcal{L}$ involves the coupling of the vector field A_μ to the conserved current: $J_\mu = i[\phi(\partial_\mu\phi^*) - \phi^*(\partial_\mu\phi)] + 2eA_\mu\phi\phi^*$, where $\phi = \frac{1}{\sqrt{2}}(\varphi_1 + i\varphi_2)$. Note also that $\mathcal{L}$ is invariant under the *local* U(1) gauge transformations:

$$\begin{aligned}\phi(x) &\to \phi'(x) = e^{-i\alpha(x)}\phi(x),\\ A_\mu(x) &\to A'_\mu(x) = A_\mu(x) + \frac{1}{e}\partial_\mu\alpha(x),\end{aligned} \tag{2}$$

where the gauge function $\alpha(x)$ is an arbitrary scalar function of the space and time coordinates.[3]

Higgs chose the potential V to have a *minimum* at the following values of the scalar fields: $\varphi_1(x) = 0$, $\varphi_2(x) = \varphi_0 = constant$; i.e., $V'(\varphi_0{}^2) = 0$, $V''(\varphi_0{}^2) > 0$, where single and double primes denote respectively the first and second derivatives with respect to scalar fields. The mathematical form of V indicates that this particular choice of the vacuum values of the scalar fields is only one out of an *infinite* number of possibilities that the potential V can acquire its minimum value. In field theoretic language, this in turn amounts to the *spontaneous* breaking of the local U(1) gauge symmetry by the chosen ground-state; in the sense that while $\mathcal{L}$ is *fully* invariant

[3] $\mathcal{L}$ is also invariant under the *global* U(1) phase transformation: $\phi \to \phi e^{i\theta}$, where θ is an arbitrary real constant.

under the local U(1) gauge transformations in (2), the chosen ground-state does *not* display this symmetry.[4]

By using the *variational principle* and taking into account only the *small* variations of the scalar fields and of the vector field, Higgs obtained the Euler-Lagrange field equations for *small* displacements of the fields around the chosen ground-state: $\varphi_1(x) = 0, \varphi_2(x) = \varphi_0$:

$$\partial^\mu \left\{ \partial_\mu \left(\Delta\varphi_1 \right) - e\varphi_0 A_\mu \right\} = 0, \tag{3a}$$

$$\left\{ \partial^2 - 4\varphi_0^2 V'' \left(\varphi_0^2 \right) \right\} \left(\Delta\varphi_2 \right) = 0, \tag{3b}$$

$$\partial_\nu F^{\mu\nu} = e\varphi_0 \left\{ \partial^\mu \left(\Delta\varphi_1 \right) - e\varphi_0 A^\mu \right\}, \tag{3c}$$

where only the linear terms have been kept, and $\Delta\varphi_1$ and $\Delta\varphi_2$ denote the small variations of the scalar fields around the ground state. Note that, after the spontaneous breaking of the local U(1) gauge symmetry by the chosen ground-state, the local U(1) gauge invariance of the theory is still maintained, in that all the local gauge transforms of the fields that are the solutions of Eqs. 3a–3c are also the solutions of the same equations. Here, since Higgs's treatment takes into account only the first order variations of the fields around the ground-state, the local U(1) gauge transformations in (2) should be considered in a *linear approximation*, in which the gauge function α is also taken to be *small*.[5]

By using the local U(1) gauge invariance of $\mathcal{L}$, Higgs introduced a new vector field B_μ through the "unitary gauge" transformation on A_μ; i.e., $B_\mu = A_\mu - (e\varphi_0)^{-1}\,\partial_\mu(\Delta\varphi_1)$, which in turn leads to $G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu = F_{\mu\nu}$. In the unitary gauge, Eqs. 3a and 3c take the following forms, respectively:

$$\partial_\mu B^\mu = 0, \tag{4a}$$

$$\partial_\nu G^{\mu\nu} + e^2 {\varphi_0}^2 B^\mu = 0. \tag{4b}$$

Higgs noted that Eqs. 4a and 4b jointly describe vector waves whose quanta have a mass of $e\varphi_0$, indicating that B_μ is a *massive* vector field. Higgs also pointed out that if there were no conserved current-vector field coupling, corresponding to the value of zero for the coupling constant e, Eqs. 3a and 3c would describe *massless* scalar bosons and *massless* vector bosons, respectively, meaning that $\Delta\varphi_1$ would be the massless Goldstone boson field. However, in the presence of the conserved current-vector field coupling, $\Delta\varphi_1$ is transformed away into the *longitudinal* polarization

[4] In the same sense, the global U(1) phase symmetry is also *spontaneously* broken by the chosen ground-state.

[5] Under the "small" local U(1) gauge transformations, $\Delta\varphi_1 \to \Delta\varphi'_1 = \Delta\varphi_1 + \alpha\varphi_0$; $\Delta\varphi_2 \to \Delta\varphi'_2 = \Delta\varphi_2$.

state of the massive vector field B_μ by means of the unitary gauge transformation. This is indicated by the presence of the term $\partial_\mu(\Delta\varphi_1)$ in B_μ; given that, in field theory, the divergence of a longitudinal wave is non-zero and its curl is zero, whereas the divergence of a transverse wave is zero and its curl is non-zero.

Hence, following Anderson's proposal, Higgs demonstrated that in a gauge theory of the Yang-Mills type spontaneous breaking of *local* gauge symmetry brought about a massive vector field without destroying local gauge invariance and without giving rise to the unwanted Goldstone bosons, leading him to conclude that "[t]his phenomenon is just the relativistic analog of the plasmon phenomenon to which Anderson [(1963)] has drawn attention" (1964, p. 508). An important consequence of the mass-generation mechanism suggested by Higgs – often referred to as the "Higgs mechanism" – is that it also brings about a *massive scalar boson* – today referred to as the "Higgs boson" – as indicated by Eq. 3b that describes scalar waves whose quanta have a mass of $2\varphi_0(V''(\varphi_0{}^2))^{1/2}$. Incidentally, the long-sought Higgs boson has been recently discovered in the ATLAS and CMS experiments currently running at CERN (see ATLAS Collaboration 2012; CMS Collaboration 2012).

Due to space limitation, I can only very briefly mention here that Englert and Brout (1964) and Guralnik et al. (1964) suggested essentially the same mechanism of mass-generation following different approaches, albeit without the prediction of a massive scalar boson (for historical details, see Karaca 2013). In passing, let me also note that recently this mass-generation mechanism has been sharply criticized in different ways by philosophers of science – see Earman 2004; Smeenk 2006; Healey 2007; Lyre 2008.

3 Characterization of Anderson's Analogy

We have seen that the analogy drawn by Nambu and Jona-Lasinio to the energy-gap structure in the BCS theory enabled the adaptation of the concept of spontaneous symmetry breaking into the context of elementary particle physics through a model of nucleon-mass generation, where the way the nucleon mass is acquired is analogous to the formation of energy-gap in the BCS theory of superconductivity. This analogy had a "heuristic" value in the construction of the Higgs mechanism in the sense that it highlighted the significance of spontaneous symmetry breaking for the solution of the mass problem in gauge theories. However, it was *primarily* aimed at solving the problem of nucleon-mass generation that is qualitatively different from the mass problem of gauge theories of the Yang-Mills type. The former is a problem of mass generation for *fermions*; whereas the latter is a problem of mass-generation for *vector bosons*; namely, the question of how vector bosons acquire mass in gauge theories of the Yang-Mills type. Therefore, as I shall characterize in what follows, it is the analogy drawn by Anderson to the explanation of the Meissner effect by the plasmon theory that enabled the construction of the Higgs mechanism.

Gentner's "structure-mapping" theory of analogy (1983) provides a sufficient ground to characterize Anderson's analogy. The basic tenet of Gentner's theory is that "an analogy is an assertion that a relational structure that normally applies in one domain [called "base"] can be applied in another domain [called "target"]" (p. 156). Therefore, according to Gentner's theory, what characterizes an analogy is the mapping of relations among individual things from a base-domain to a target-domain, rather than the mapping of attributes of things. Applying Gentner's theory to the case of Anderson's analogy, we can identify the relata of the base-relation as massless phonons in a superconductor and photons which are the massless vector bosons of the electromagnetic interaction; the base-relation being that, as a consequence of conserved current (i.e., Cooper pair current)-vector field (i.e., external magnetic field) coupling, massless phonons, which have only one longitudinal polarization state, combine with massless photons, which have only two transverse polarization states, to form massive plasmons which have one longitudinal and two transverse polarization states. Thus, Anderson's analogy can be characterized by the mapping of this base-relation from the context of the plasmon theory of solid-state physics – i.e., base-domain – to the context of the gauge theories of the Yang-Mills type – i.e., target-domain.

In Anderson's analogy, the relata of the target-relation can be identified as the Goldstone field resulting from the spontaneous breaking of local gauge symmetry and the massless Yang-Mills vector field; the target relation being that as a consequence of conserved current-vector field coupling the Goldstone field, which has only one longitudinal polarization state, combines with the massless Yang-Mills vector field, which has only two transverse polarization states, to form the massive Yang-Mills vector field which has one longitudinal and two transverse polarization states.

In Anderson's analogy, the base relation is a description of a mechanism that accounts for the Meissner effect – referred to as the "Anderson mechanism" in solid-state physics. Anderson's analogy suggests the target-relation to be a description of a mechanism – namely, the Higgs mechanism – that accounts for the way vector bosons acquire mass in gauge theories of the Yang-Mills type. As we have previously seen, the target relation suggested by Anderson's analogy was later adapted to field theory by Higgs as a solution of the mass problem of the Yang-Mills theory. Given that the Higgs mechanism is a constitutive element of the Glashow-Weinberg-Salam electroweak theory, this discussion also illustrates the role of analogy in theory construction; a topic that was extensively discussed in the literature of philosophy of science, prominently by Hesse (1966).

4 Conclusions

In this paper, I have argued that Anderson's analogy to the treatment of the Meissner effect by the plasmon theory of solid-state physics provided key guidance to Higgs as to how to account for the mass problem of gauge theories of the Yang-Mills type.

As I have shown, Anderson's analogical proposal indicated not only the relation that was required to hold between the Goldstone and massless vector fields in order for the latter to acquire mass, but also the condition, namely, spontaneous breaking of *local* gauge symmetry, under which this underlying relation could hold in gauge theories of the Yang-Mills type. Higgs demonstrated the validity of Anderson's proposal in a gauge theory of the Yang-Mills type displaying spontaneous breaking of *local* U(1) gauge symmetry, and he showed that the relation previously suggested by Anderson held between the Goldstone fields and the massless vector fields and that thereby the latter acquired mass in a way that did not destroy the local gauge invariance of the theory.

These considerations indicate that Anderson's analogical argument served to guide the construction of the Higgs mechanism, suggesting the "heuristic dependence" of elementary particle physics on solid-state physics – in Grantham's sense. Remember that, in Grantham's account, one of the ways in which the practical unification of two fields is obtained is through the heuristic dependence of one on the other for the construction of a novel hypothesis. Therefore, I conclude that the construction of the Higgs mechanism was achieved through the practical unification of solid-state and elementary particle physics – in Grantham's sense. Note that even though Grantham takes "heuristic dependence" to be one of the ways in which practical unification of fields can take place, he does not tell us enough about how such dependence between fields can actually happen in scientific practice. The present case study shows that the heuristic dependence of one field on the other can be established through analogies drawn between fields.

The present case study also shows that, in addition to enabling the construction of the Higgs mechanism, the practical unification of solid-state and elementary particle physics also contributed, albeit indirectly, to the construction of the electroweak theory; in that Weinberg (1967) used the Higgs mechanism to construct this theory (see Karaca 2013). This illustrates how the practical unification of fields can enable the construction of a novel theory in a scientific discipline, thereby lending *operational* meaning to the notion of practical unification insofar as it provides a methodology of theory construction in the practice of modern physics.

Due to the prevalence of the accounts of theoretical unification in philosophical literature, unity in science has been often viewed as originating solely from the formal relations involved in the structures of theories, thus reflecting merely the formal and abstract aspects of the process of hypothesis/theory construction in science (for the discussion of theoretical unification in the case of the electroweak theory; see, e.g., Maudlin 1996; Wayne 1996; Morrison 2000). By contrast, the practical unification of solid-state and elementary particle physics suggested in the present paper is *contextual* rather than formal and abstract; in that it primarily concerns physicists' practice of drawing analogies to the theories in their neighboring fields and using them to solve conceptual problems as well as to construct novel theories and theoretical mechanisms in their own fields.

References

Anderson, P. W. (1963). Plasmons, gauge invariance, and mass. *Physical Review, 130*, 439–442.

ATLAS Collaboration (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B, 716*, 1–29.

Bardeen, J., Cooper, L. N., & Schrieffer, J. R. (1957). Theory of superconductivity. *Physical Review, 108*, 1175–1204.

Brown, L. M., & Cao, T. Y. (1991). Spontaneous breakdown of symmetry: Its rediscovery and integration into quantum field theory. *Historical Studies in the Physical and Biological Sciences, 21*, 211–235.

Cat, J. (1998). The physicists' debates on unification in physics at the end of the 20th century. *Historical Studies in the Physical and Biological Sciences, 28*(part 2), 253–299.

Cat, J. (2006). Fuzzy empiricism and fuzzy-set causality: What is all the fuzz about? *Philosophy of Science, 73*, 26–41.

Cat, J. (2007). The unity of science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/scientific-unity/

CMS Collaboration (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B, 716*, 30–61.

Darden, L., & Maull, L. (1977). Interfield theories. *Philosophy of Science, 44*, 43–64.

Earman, J. (2004). Laws, symmetry, and symmetry breaking: Invariance, conservation principles, and objectivity. *Philosophy of Science, 71*, 1227–1241.

Englert, F., & Brout, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Physical Review Letters, 13*, 321–323.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170.

Goldstone, J. (1961). Field theories with superconductor solutions. *Nuovo Cimento, 19*, 154–164.

Goldstone, J., Salam, A., & Weinberg, S. (1962). Broken symmetries. *Physical Review, 127*, 965–970.

Grantham, T. (2004). Conceptualizing the (dis)unity of science. *Philosophy of Science, 71*, 133–155.

Guralnik, G. S., Hagen, C. R., & Kibble, T. W. B. (1964). Global conservation laws and massless particles. *Physical Review Letters, 13*, 585–587.

Healey, R. (2007). *Gauging what's real: The conceptual foundations of contemporary gauge theories*. Oxford: Oxford University Press.

Hesse, M. B. (1966). *Models and analogies in science*. Notre Dame: University of Notre Dame Press.

Higgs, P. W. (1964). Broken symmetries and masses of gauge bosons. *Physical Review Letters, 13*, 508–509.

Karaca, K. (2013). The construction of the Higgs mechanism and the emergence of the electroweak theory. *Studies in History and Philosophy of Modern Physics, 44*, 1–16.

Lyre, H. (2008). Does the Higgs mechanism exist? *International Studies in the Philosophy of Science, 22*, 119–133.

Maudlin, T. (1996). On the unification of physics. *Journal of Philosophy, 93*, 129–144.

Morrison, M. (2000). *Unifying scientific theories: Physical concepts and mathematical structures*. Cambridge: Cambridge University Press.

Nambu, Y. (1960). Quasi-particles and gauge invariance in the theory of superconductivity. *Physical Review, 117*, 648–663.

Nambu, Y., & Jona-Lasinio, G. (1961). Dynamical model of elementary particles based on an analogy with superconductivity I. *Physical Review, 122*, 345–358.

Nozieres, P., & Pines, D. (1958). Electron interaction in solids. General formulation. *Physical Review, 109*, 741–761.

Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science* (Vol. 2, pp. 3–36). Minneapolis: Minnesota University Press.

Peskin, M. E., & Schroeder, D. V. (1995). *An introduction to quantum field theory*. Reading: Addison-Wesley.

Schwinger, J. (1962). Gauge invariance and mass. *Physical Review, 125*, 397–398.

Smeenk, C. (2006). The elusive Higgs mechanism. *Philosophy of Science, 73*, 487–499.

Wayne, A. (1996). Theoretical unity: The case of the standard model. *Perspectives on Science, 4*, 391–407.

Weinberg, S. (1967). A model of leptons. *Physical Review Letters, 19*, 1264–1266.

Part VI
Philosophy of the Physical Sciences: Philosophy of Space and Time

A Critical Note on Time in the Multiverse

Svend E. Rugh and Henrik Zinkernagel

Abstract In recent analyses of standard, single-universe, cosmology, it was pointed out that specific assumptions regarding the distribution and motion of matter must be made in order to set up the cosmological standard model with a global time parameter. Relying on these results, we critically examine the notion of time in the multiverse – and in particular the idea that some parts of the multiverse are older than others. By focusing on the most elaborated multiverse proposal in cosmology, the inflationary multiverse, we identify three problems for establishing a physically well-defined notion of global time; a quantum problem, a collision problem and a fractal problem. The quantum problem – and the closely related "cosmic measurement problem" – may even undermine the idea that parts of the multiverse causally and temporally precede our universe.

1 Introduction

The idea of a multiverse has recently become quite popular in modern cosmology. According to some multiverse scenarios, based e.g. on so-called chaotic inflation, our universe is supposed to be just one inflating bubble in a much bigger and older multiverse with each component expanding differently and having different physical laws (see e.g. Linde 2004; Guth 2007). In this and related versions, the multiverse thus purports to reject the common wisdom regarding modern cosmology according

S.E. Rugh
Symposion, 'The Socrates Spirit', Section for Philosophy and the Foundations of Physics,
Hellebækgade 27, Copenhagen N, Denmark
e-mail: rugh@symposion.dk

H. Zinkernagel (✉)
Department of Philosophy I, Granada University, 18071 Granada, Spain
e-mail: zink@ugr.es

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_22,

to which asking what was before the big bang is considered as meaningless as asking what is north of the North Pole, see e.g. Hawking (1989, p. 69).

While the multiverse idea has been widely discussed (and criticized) e.g. in connection with its apparent lack of empirical testability (see e.g. Carr and Ellis 2008) very few studies have addressed the more conceptual problems facing the notion of a multiverse in cosmology.[1] In this paper, we want to explore a little discussed conceptual question about the multiverse: Does it include a sensible *notion of time* which allows us to speculate that it is not only much bigger but also much *older* than our (local) universe? The answer to this question will obviously depend both on what kind of multiverse is contemplated, and on how time is (or could be) conceived in the specific multiverse proposal. In any case, the investigation of the question is likely to contribute to a clarification of the conceptual foundation of cosmology.

The outline of the paper is as follows. We first review some earlier work which shows that a relationist understanding of time (an interdependence between time, matter and motion) is essential to the standard notion of cosmic time. Armed with this clarification, we discuss possible ways to understand the claim that there are older patches (than our universe) in the multiverse. After that we discuss the most worked out version of the multiverse arising from the theory of inflation and question whether the notion of time in this theory is applicable as a multiverse time. In the closing section, we offer a few brief comments on other multiverse scenarios and note that these are likely to be even worse off with regard to time than the inflationary multiverse.[2]

2 Time in Standard (Single-Universe) Cosmology

In our earlier work we have defended a version of relationism which affirms that time is necessarily associated with physical processes. More specifically, we argue in favour of a 'time-clock' relation which asserts that time, in order to have a physical basis, must be understood in relation to physical processes which act as 'cores' of clocks (Rugh and Zinkernagel 2005, 2009, see also Zinkernagel 2008). In the cosmological context, the time-clock relation implies that a necessary physical condition for *interpreting* the t parameter of the standard Friedmann-Lemaître-Robertson-Walker (FLRW) model as cosmic time in some 'epoch' of the universe is the (at least possible) existence of a physical process which can function as a core of a clock in the 'epoch' in question.

There is a more direct route to relationism in cosmology which is independent of the mentioned time-clock-relation (even if in conformity with it). In this regard, we

[1] For a brief review and references to some of these problems, see Zinkernagel (2011).

[2] We explore the notion of time in both the universe and the multiverse in more detail in Rugh and Zinkernagel (2014).

discuss in Rugh and Zinkernagel (2011) how the very set-up of the FLRW model with a global time is closely linked to the motion, distribution and properties of cosmic matter. In the following, we briefly review some key points of this discussion which are necessary components of our analysis of time in the multiverse.

In relativity theory time depends on the choice of reference frame. Since, for a universe, a reference frame cannot be given from the outside, such a frame has to be "built up from within", that is, in terms of the (material) constituents within the universe. It is often assumed that the FLRW model may be derived just from the cosmological principle. This principle states that the universe is spatially homogeneous and isotropic (on large scales). It is much less well known that another assumption, called Weyl's principle, is necessary in order to arrive at the FLRW model and, in particular, its cosmic time parameter. Whereas the cosmological principle imposes constraints on the *distribution* of the matter content of the universe, Weyl's principle imposes constraints on the *motion* of the matter content. Weyl's principle (from 1923) says that the matter content is so *well behaved* that a reference frame can be built up from it:

> Weyl's principle (in a general form): The world lines of 'fundamental particles' form a spacetime-filling family of non-intersecting geodesics (a congruence of geodesic world lines).

The importance of Weyl's principle is that it provides a reference frame which is physically based on an expanding 'substratum' of 'fundamental particles' (e.g. galaxies or clusters of galaxies). In particular, if the (non-crossing) geodesic world lines are required to be orthogonal to a series of space-like hypersurfaces, a comoving reference frame is defined in which constant spatial coordinates are "carried by" the fundamental particles (see Fig. 1 in Sect. 3.1). The time coordinate is a cosmic time which labels the series of hypersurfaces, and which may be taken as the proper time along any of the particle world lines. We note that the congruence of world lines is essential to the standard cosmological model since the symmetry constraints of homogeneity and isotropy are imposed w.r.t. such a congruence (see e.g. Ellis 1999). Thus, Weyl's principle is *a precondition* for the cosmological principle; the former can be satisfied without the latter being satisfied but not vice versa.

2.1 Is the Weyl Principle (Always) Satisfied in Our Universe?

There are several possible problems which may arise with the Weyl principle. First, there is the question of whether particle trajectories are always well-defined (at all times in cosmic history). Second, whether – if such well-defined trajectories cross – a suitable averaging procedure exists for smoothing out these crossings. As regards the latter problem, it is clear that Weyl's principle cannot hold for ordinary galaxies as they indeed may (and do) collide. Likewise with the more fundamental constituents in earlier phases of the universe. Thus the fundamental world lines in the Weyl principle must be some ‘average world lines’ associated with the average motion of the fundamental particles over some coarse-grained scales (in order to “smooth out” any crossings).[3]

Regarding the first problem of whether particle trajectories can at all be identified, the starting point is that the Weyl principle refers to a non-crossing family of (fluid or particle) world lines. The notion of such lines refers to classical, or classicalized, particle-like behavior of the material constituents. This makes it difficult to even formulate the Weyl principle (let alone decide whether it is satisfied) if some period in cosmic history is reached (in a backward extrapolation from now) where the ‘fundamental particles’ are to be described by wave-functions $\psi(x,t)$ referring to entangled quantum constituents. What is a ‘world line’ or a ‘particle trajectory’ then? Unless one can specify a clear meaning of non-intersecting trajectories in a contemplated quantum ‘epoch’, it would seem that the very notion of cosmic time, and hence the notion of ‘very early universe’ is compromised. This last problem of identifying a Weyl substratum within a quantum description arises most clearly on a “quantum fundamentalist” view according to which the material constituents of the universe could be described *exclusively* in terms of quantum theory at some early stage of the universe. As noted in Rugh and Zinkernagel (2011), there is still no good answer to what may be called the “cosmic measurement problem” (how to get classical structures from quantum constituents in a cosmological context), not least because it is highly questionable whether decoherence is sufficient to explain the building up of a Weyl substratum.

3 Time in the Multiverse?

With the above considerations concerning time in standard cosmology, we are now ready to tackle the question of time in the multiverse. More specifically, we ask whether – and under which conditions – one is justified in contemplating the idea that some parts of the multiverse are older than ours. There seem to be at least two relevant ways to establish the possibility of older patches or bubbles:

1. Define some sort of a ‘multiverse’ (or ‘supercosmic’) time for the multiverse which gives a definite time ordering of the patches (as in Fig. 2 below).

[3] There exist observationally based claims (e.g. Labini et al. 2009) that the matter distribution is not homogeneous but instead fractal at intermediate scales at least up to distances of the order $\sim$100 Mpc. If this fractality extended to arbitrary large distance scales, there would be no scale above which collisions could be averaged out. Moreover, there would be ‘holes’ on all scales so no set of ‘average world lines’ could fill space-time (implying that no congruence could be formed), and also a homogeneous universe could not be recovered. Thus, both the Weyl principle and the cosmological principle – even in their ‘coarse grained’ versions – would be undermined (see Rugh and Zinkernagel 2014).

2. If this cannot be done, then try to extrapolate our ‘local’ cosmic time concept back through our ‘local’ big bang.[4]

Either way, the overall conclusion from Sect. 2 is that time is relational. Thus, there is no freely flowing absolute and universal background time parameter so both multiverse – and (the extrapolation of a) cosmic time need to be grounded in the behavior of the constituents within the multiverse and the universe respectively.

The notion of a ‘multiverse’ covers a great many possibilities (see e.g. Carr 2007). In order to address something relatively well-defined we shall in this short note restrict ourselves to consider some particular case studies of inflationary multiverse models which, in our assessment, seem to be models (1) in which the model builders to some degree reflect upon – or even attempt to provide a physical underpinning for – the time concepts employed; and (2) which are investigated and developed to a degree that they have entered the contemporary standard literature on cosmology with some claims of observational testability.[5] The basic idea of the inflationary multiverse is that of a background (inflating) de Sitter space in which local bubble universes (where inflation quickly comes to an end in thermalization and particle production) continuously form (see Fig. 2). In its simplest version, the inflationary multiverse is driven by a single scalar field φ – the inflaton (which, at present, is unrelated to any known particle physics), see e.g. Linde (2004).

3.1 Can a Multiverse Time Be Defined?

In this paper a main question concerns whether the Weyl principle is satisfied in the multiverse. To motivate an initial doubt, consider Fig. 2, in which there does not seem to be a multiverse with patches or bubbles obeying the Weyl principle (a similar figure suggesting a multiverse time can be found in Guth 2007). Thus, there is no immediate physical basis for a multiverse time (indicated in the figure) which could order the patches.[6]

By making the analogy between the idealized Weyl substratum (a congruence of e.g. galaxy world-lines) in our universe (Fig. 1) and the picture of an infiltrated network of bubbles in the realm of the inflationary multiverse (Fig. 2), it is assumed that the bubble universes somehow play the role of the substratum. The only

[4] A third and related possibility, which we shall discuss below, is to use proper time – or at least a time order – (associated with a single world line) to extrapolate backwards even in cases where no ‘local’ cosmic time can be defined.

[5] Note that there are other proposals for older structures than our universe – e.g. cyclic universes (“temporal” multiverses). Some of these have been discussed (and their time concept criticized) in Zinkernagel (2008).

[6] The colours in the (original version of the) figure represent different effective physical laws (or constants) in the different bubble universes. This corresponds to a more complicated multiverse model – with various scalar fields – than the one discussed below (which has only one scalar field φ). In our view, however, this complication does not change the discussion to follow.

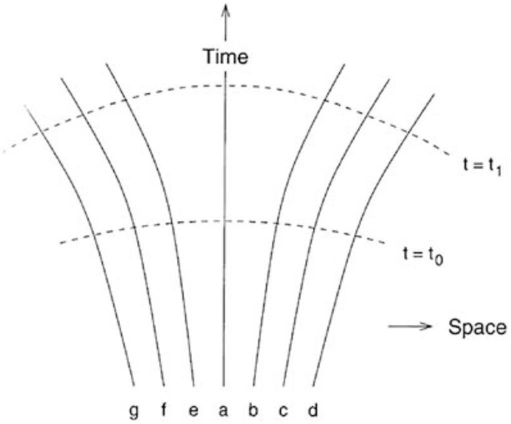


Fig. 1 An idealized "Weyl substratum". The particle (e.g. galaxy) trajectories form a congruence in an approximation where galaxies are seen as space-time filling particles of a fluid (Figure from Narlikar 2002)

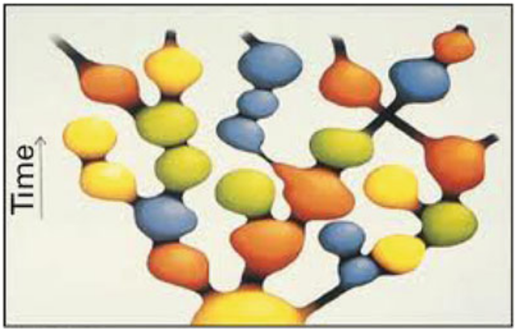


Fig. 2 A multiverse consisting of bubble universes arising from the chaotic inflation model with a suggestive global multiverse "time" axis indicated on the left (Figure from Linde 1998)

alternative will be to assume that this substratum is constructed from the part of the φ-field outside the bubbles. Either way, we see trouble. In the first case, because the bubbles collide. In the second case because it is hard to construct (Weyl) trajectories from the φ-field, e.g. due to quantum effects (see below).

The need to satisfy the Weyl principle does seem to be recognized in the multiverse literature. Thus, Vanchurin et al. (2000) notes that "an inflating universe can be locally [within a bubble] described using the synchronous coordinates $ds^2 = d\tau^2 - a^2(\mathbf{x}, \tau)d\mathbf{x}^2$". They continue:

> The lines of $\mathbf{x}$ = const in this [synchronous] metric are timelike geodesics corresponding to the world lines of co-moving observers, and the coordinate system is well defined as long as the geodesics do not cross. This will start happening only after thermalization, when matter in some regions will start collapsing as a result of gravitational instability. Hence, the synchronous coordinates can be extended to the future well into the thermalized region.

This amounts to the claim that there is a Weyl substratum (a φ-field) which allows us to set up a synchronous coordinate system within a bubble (and it is within a bubble that thermalization and particle production occur). A similar construction seems to be applied in the global case, i.e. for the whole multiverse. Thus, Guth (2007, p. 6820) remarks that one can construct "a Robertson-Walker coordinate system while the model universe is still in the false vacuum (de Sitter) phase, before any pocket universes have formed. One can then propagate this coordinate system forward with a synchronous gauge condition". However, even if the importance of the Weyl principle is implicitly recognized by proponents of the inflationary multiverse, we see three step-wise related problems for this principle to be satisfied (and hence for a multiverse time to be physically underpinned):

1. Are there well-defined trajectories in the multiverse?
2. If there are well-defined trajectories, are they non-crossing?
3. If they cross, can such crossing trajectories be "averaged out"?

We elaborate a bit on these questions in the following three subsections, and as we shall see, the answers to them may well be no, no and no. This is due to what one could call, respectively, the quantum problem, the collision problem and the fractal problem.

3.1.1 Are There Well-Defined Trajectories in the Multiverse?

As mentioned in Sect. 2.1, the assumption of a quantum nature of the material (or otherwise) constituents of the universe makes it hard (or impossible) to associate these with well-defined particle trajectories. And during inflation the only relevant constituent of the universe is taken to be the inflaton field φ which – in the last analysis – is a quantum field. While the quantum-classical transition from quantum fluctuations to classical density perturbations has been widely discussed (even if not critically scrutinized, for an exception see e.g. Sudarsky 2011), this point – that the φ field itself is a quantum field – is easily overlooked. For instance Linde writes after describing the basic mechanism in chaotic inflation (the most simple inflation model) which ends in the oscillations of the scalar field near the minimum of its potential (p. 130 in Carr 2007): "As any rapidly oscillating *classical* field, it loses its energy by creating pairs of elementary particles" (our emphasis). Despite the wording, this is not a reconceptualization of the whole edifice of classical field theory! Linde is, of course, well aware that it requires quantum fields to create particles, and that the word 'classical' simply refers to the lowest order approximation in quantum field theory. But, again, just like wave functions in non-relativistic quantum theory do not give rise to physical motion (of a particle or wave) in space and time – without assumptions solving the measurement problem – so quantum fields do not describe moving elementary particles in space with well-defined trajectories.

If we assume that this 'quantum problem' could be properly dealt with, we would then have a sufficiently classical (or classicalized) inflaton field φ. The existence of such a field has been assumed (as a mere postulate) in the investigation of various scalar field inflationary models since their inception in the early 1980s. The background space for the inflationary multiverse is de Sitter space in which no matter is present (matter is only produced at the end of inflation inside the bubbles). Thus, the multiverse has – as available ‘material’ to build up the reference frame from within – only the inflaton field φ. From this inflaton field one should construct some trajectories in order to satisfy the Weyl principle and thereby provide multiverse time (de Sitter t) a physical underpinning.[7] One way of getting (a congruence of) non-crossing trajectories is to assume that the matter-energy content is in the form of a perfect fluid since this implies a well-defined four-velocity (and hence a direction for a trajectory) at each point of the spacetime manifold. As described e.g. by Krasinski (1997, p. 8) and Hobson et al. (2006, p. 432), a 4-velocity field of a perfect fluid can be constructed from (the gradient of) a scalar field.[8] However, as we shall see in Sect. 3.1.3 below, this may not be sufficient to satisfy the Weyl principle due to the fractal structure of the inflationary multiverse.

3.1.2 Could the Trajectories Be Non-crossing?

If the relevant substratum for the Weyl principle in the (inflationary) multiverse is the bubble- or pocket universes, there does indeed seem to be crossing of the trajectories.[9] For instance, Garriga, Guth and Vilenkin (2007) note:

> A bubble universe nucleating in an eternally inflating false vacuum will experience, in the course of its expansion, collisions with an infinite number of other bubbles.

Thus, bubble collisions do occur and so the Weyl principle is not satisfied at the level of bubbles. This problem appears to be aggravated by the observation that the inflationary multiverse seems to result in a fractal structure in which merging of different thermalized domains (bubble universes) occurs on all scales (see e.g. Guth 2007 and Vanchurin et al. 2000).

[7] Greene (2012, p. 69) suggests that one may directly use the changing value of the φ field as a clock (as measured by an “inflaton-meter”). He apparently assumes that the φ value is monotonically decreasing in de Sitter t. This idea is similar to the standard use, in FLRW cosmology, of matter density ρ, or the temperature T of the background radiation, as a clock, see e.g. discussions in Rugh and Zinkernagel (2009). However, in our assessment, Greene’s clock cannot in general trace the de Sitter t “time” parameter (and thus cannot provide a physical underpinning of it). First because the (classical part of the) φ field may not be homogeneous in x-space (as in Linde’s chaotic model) – and so the same φ value (an ‘equal φ hypersurface’) becomes associated with different de Sitter t-values. And second because even an assumed homogeneously distributed φ field will exhibit quantum fluctuations so that, again, the same φ value gets mixed up with different t-values.

[8] The idea is to equate the energy momentum tensor of the perfect fluid form with the energy momentum tensor for the scalar field. This results in the 4-velocity $u_\mu = A \cdot \partial_\mu \varphi$ where $A = (\partial^\nu \varphi \; \partial_\nu \varphi)^{-1/2}$.

[9] If the substratum (the world-lines of which “carry” the coordinates) is not formed by the bubble universes, but is rather to be found in the background de Sitter space with an inflaton field, then we are either back in the subsection above or proceed to the subsection below.

3.1.3 Could Crossing Trajectories Be "Averaged Out"?

Even if bubbles collide, and so trajectories cross, it may still be possible – just as in the single universe case – to devise an averaging procedure to "smooth out" these crossings. However, this will be difficult in the realm of the inflationary multiverse since it appears to be fractal (Guth 2007, p. 6816). This means, as far as we can see, that there is no "cut off" scale above which the implementation of averaging procedures will produce non-colliding world-line trajectories out of bubbles (which collide below such a scale).

If the Weyl substratum is to be constructed from the φ field (outside the bubbles) the situation seems no better since these regions outside the bubbles likewise appear to form a fractal. This is suggested e.g. by the highly random and irregularly looking distribution of the scalar field(s) in Fig. 20.2 in Linde (2004, p. 435) and explicitly stated in Vanchurin et al. (2000, p. 4): "…these [inflating] regions [outside the bubbles] form a fractal of dimension $d < 3$". Although this may not result in collisions between trajectories constructed from the φ field, it nevertheless seems to imply a problem concerning the averaging procedure. According to Guth (2007, p. 6816),

> One does have to think about the fractal structure if one wants to understand the very large scale structure of the spacetime produced by inflation.

We agree. But if one, indeed, thinks about exactly this, it appears that the fractal structure of the inflationary multiverse results in a far more complicated large scale spacetime structure than the highly symmetric Robertson-Walker spaces (which are isotropic and homogeneous) employed in simplified inflationary modeling. More fundamentally, in our assessment (and to be examined further), the Weyl principle appears not to be satisfied: According to this principle, the reference frame is built up from a space-time filling congruence of geodesics. This can at most be fulfilled in a coarse grained (averaging) sense. However, due to the self-similar fractal structure (of both the inflating and thermalized – bubble – regions) there is no possible coarse graining scale above which a spacetime filling congruence can be constructed (as there will be 'holes' at all scales). If this is so, the physical foundation for a global de Sitter multiverse time appears insufficient.

3.2 Extrapolating Our Cosmic Time, Proper Time or Time Order Back to an Older Bubble?

If the Weyl principle does not hold in the multiverse, there will be no global time parameter which can be used to temporally order the bubble universes of *different* 'branches' in Fig. 2. But it would seem that, even without a Weyl principle, it should still be possible to contemplate older structures than our universe by focusing on a *single* (our own) 'branch' in the figure. Indeed, if we accept the idea that one bubble universe can somehow causally give rise to another, then it appears possible to consider other bubble universes (within our own ‘causal branch’) which predate our universe. Nevertheless, as we shall indicate below, to contemplate this possibility may be far from straightforward.

One way to address the causal past of our universe would be if we could extrapolate our ‘local’ cosmic time concept further back than our (local) beginning. Now, if this beginning is taken to be (arbitrarily close to) an initial singularity or, alternatively, that it is located in some ‘epoch’ described by quantum gravity such a proposal seems hopeless or, at best, highly speculative (see also Zinkernagel 2008). Indeed, most cosmologists would agree that there is no (known) sensible time concept “before” the Planck time ($\sim 10^{-43}$ s) and so no clear meaning can be ascribed to instants earlier than that.

However, if the beginning of our universe occurs – as assumed in inflationary multiverse models – at the beginning of the inflationary phase, then there may be no need to extrapolate time either through a singularity or through a quantum gravity epoch. Indeed, as long as some causal structure can be maintained (light cones should not tilt more than 45°), then it may be sensible to speak of the past of any event. Thus, one may perhaps speculate, for instance, that before the beginning of inflation at, say 10^{-35} s, the universe no longer gets denser and hotter (as in standard cosmology) but rather expands into a previous bubble universe. In fact, such a suggestion may work even if the ‘local’ (in our universe) Weyl principle is not satisfied in the inflationary epoch. For even if there is no cosmic time (no Weyl principle) it could still be possible to ask about the past of any event – for instance, the past of the onset of inflation. Specifically, we can address the past of an event by extrapolating backwards proper time along a world-line which ends in the event. Such a possibility appears to be implied when Tegmark (2005, p. 49) remarks (after stating, as we saw Garriga, Guth and Vilenkin do above, that geodesics cross after thermalization within a bubble):

> When we discuss t [time] for a particle in the present epoch, the rigorously inclined reader can simply take this to mean its proper time, since this provides a well-defined ordering even after geodesic crossing. [our inserts]

For this to be made into a workable suggestion for contemplating earlier bubbles than our own, it must be possible to identify (or, at least, to speculate) a particle world-line along which proper time can be extrapolated backwards.[10] In particular, photons – or other massless particles – alone will not be sufficient as they have no past (i.e. their proper time is zero).[11] Note that proper time along a specific

[10] From our relationist point of view – in which time is necessarily related to physical processes (Rugh and Zinkernagel 2009) – the time-like curves can only be identified (they only have a physical basis) if the motion of objects or test particles along these curves is at least in principle realizable from the available physics.

[11] Within the framework of general relativity the notion of “causal order” depends on the construction of “backwards light cones” based on the existence of time-like or null-like curves (see e.g. Hawking and Ellis 1973, Sect. 6 “Causal Structure”) – and therefore on the notion of (possible) classical particle or light-signal trajectories. The latter is insufficient to establish a chronological

world-line will give a quantitative measure of time differences between events. But since we are here only interested in the notion of *earlier* bubbles, a time (or chronological) order will be sufficient. Thus, the existence of *any* time-like curve (on which we can address proper times $\tau < \tau_0$, where τ_0 is the beginning of our bubble) will suffice.

In the inflationary scenario, the relevant candidate for a particle world-line (a time-like curve) will have to come from the φ field. However, as discussed in Sects. 2.1 and 3.1.1, there is a 'quantum problem' in constructing sensible notions of particle world-lines and classical trajectories from the inflaton field. In particular, at the supposed 'birth' of a new bubble universe, the inflaton field is strongly quantum: Quantum fluctuations with amplitudes (within a factor of 10) of the order of the Planck scale are necessary to reset or lift the scalar field back to a value where a new bubble is born and becomes dominated by inflation (see e.g. Linde 2004, Sect. 4).[12] Thus, at the 'birth' of a new bubble universe, the φ field is nowhere close to being a classical field on top of which we have small quantum fluctuations. Rather, it is entirely dominated by Planck scale quantum fluctuations.

It is therefore unclear to us how one would go about constructing any individual classical particle world-line from the inflationary scalar field φ in a regime where its quantum behaviour is dominant. But if such world-lines (classical trajectories) cannot be constructed from the underlying physics (the φ field), it seems, in our assessment, that the very conditions for speaking about the past of an event in general relativity are not fulfilled. We therefore tentatively conclude that this proper time, or time order, route to contemplating earlier patches or bubbles (within a given branch of bubbles) in the multiverse seems problematic.[13]

4 Outlook

In this note we have argued that it is very difficult to construct a global multiverse time parameter (as suggested e.g. by Linde and Guth) which would give a temporal ordering of different branches in the inflationary model of the multiverse (cf. Fig. 2).

ordering of bubbles since – if only light is present – causal influences are instantaneous (again, photons have no past).

[12] Whereas Linde (2004) mostly discusses chaotic inflation, the quantum problem also shows up in the multiverse model based on the "new inflation" scenario: It is hinted e.g. in Vilenkin (2004) that within new inflation, the scalar field is dominated by its quantum behavior when new bubble universes form (near the maximum of the inflaton potential).

[13] Guth (2007, p. 6822) reports a theorem according to which eternal inflation is not past-eternal (i.e. there must be a beginning of the inflating multiverse even though inflation always continues somewhere). This theorem focuses on the idea of a time-like (or null-like) geodesic which is, locally, extracted backwards to an ultimate (for the multiverse as a whole) big bang. The theorem seems to rest on the idea of a well-defined 'local' congruence (of massive test particle world-lines) intersecting the geodesic. We would, again, object that the definitions of both the geodesic trajectory and the congruence are suspect if the underlying theory is of a quantum nature.

We have also indicated that it is not straightforward to maintain even a concept of time order within a given branch of bubbles since, at the birth of a bubble, the physics is entirely dominated by quantum fluctuations. This means that there is no possibility to construct classical trajectories (from the inflaton field) on which the causal and temporal structure in general relativity is based. Thus, it is difficult to provide a physical underpinning of what one could mean by saying that some other bubble universe predates our own.

Our discussion above applies only to the restricted class of (inflationary) multiverse models considered. As noted, these models appear to be the most elaborated versions of the multiverse – in particular in terms of contemplated spatio-temporal structure (e.g. the notion of a background de Sitter space). In any case, it seems to us that it might even be more problematic to think of patches or bubbles 'older' than ours if we consider more radical versions of the multiverse (for instance those contemplated in Tegmark's (2004) level III–IV). Such versions may include the notion of completely disconnected regions and/or fundamentally different physical laws in the different bubbles. This may well undermine (1) the causal structure needed to define the past light-cone of an event and, in particular, the idea of extrapolating proper time backwards to an earlier bubble; and (2) the possibility of comparing the time concepts of – and thus temporally order – different bubbles (e.g. since, as discussed in Rugh and Zinkernagel 2009, time is implicitly defined by laws). None of this means that there could not be ways to contemplate a multiverse older than our universe. But we would at least recommend that multiverse model builders ought to be clear about what time concept they use.

Acknowledgements We would like to thank Robert Brandenberger, George Ellis, Holger Bech Nielsen and an anonymous referee for helpful comments. We gratefully acknowledge the hospitality provided by the Niels Bohr Archive under the auspices of Finn Aaserud. HZ thanks the Spanish Ministry of Science and Innovation (Project FFI2011-29834-C03-02) and the Spanish Ministry of Education (Project PR2011-0150) for financial support.

References

Carr, B. (Ed.). (2007). *Universe or multiverse?* Cambridge: Cambridge University Press.

Carr, B., & Ellis, G. F. R. (2008). Universe or multiverse? *Astronomy & Geophysics, 49*(2), 2.29–2.37.

Ellis, G. F. R. (1999). 83 years of general relativity and cosmology: Progress and problems. *Classical and Quantum Gravity, 16*, A37–A75.

Garriga, J., Guth, A. H. & Vilenkin, A. (2007). Eternal inflation, bubble collisions, and the persistence of memory. *Physical Review D*, 76, 123512, 12 pp.

Greene, B. (2012) *The hidden reality. Parallel universes and the deep laws of the cosmos*. London: Penguin Books.

Guth, A. H. (2007). Eternal inflation and its implications. *Journal of Physics A: Mathematical and Theoretical, 40*, 6811–6826.

Hawking, S. W. (1989). The edge of spacetime. In P. Davies (Ed.), *The new physics* (pp. 61–69). Cambridge: Cambridge University Press.

Hawking, S. W., & Ellis, G. F. R. (1973). *The large scale structure of space-time*. Cambridge: Cambridge University Press.

Hobson, M. P., Efstathiou, G. P., & Lasenby, A. N. (2006). *General relativity*. Cambridge: Cambridge University Press.

Krasinski, A. (1997). *Inhomogeneous cosmological models*. Cambridge: Cambridge University Press.

Labini, F. S., Vasilyev, N. L., Pietronero, L., & Baryshev, Y. V. (2009). Absence of self-averaging and of homogeneity in the large-scale galaxy distribution. *Europhysics Letters, 86*, 1–6.

Linde, A. (1998). The self-reproducing inflationary universe. *Scientific American, 9*(20), 98–104.

Linde, A. (2004). Inflation, quantum cosmology, and the anthropic principle. In J. D. Barrow, P. C. W. Davies, & C. L. Harper (Eds.), *Science and ultimate reality* (pp. 426–458). Cambridge: Cambridge University Press.

Narlikar, J. (2002). *An introduction to cosmology* (3rd ed.). Cambridge: Cambridge University Press.

Rugh, S. E., & Zinkernagel, H. (2005). Cosmology and the meaning of time (76 pp.) Distributed manuscript.

Rugh, S. E., & Zinkernagel, H. (2009). On the physical basis of cosmic time. *Studies in History and Philosophy of Modern Physics, 40*, 1–19.

Rugh, S. E., & Zinkernagel, H. (2011). Weyl's principle, cosmic time and quantum fundamentalism. In D. Dieks et al. (Eds.), *Explanation, prediction and confirmation*. The philosophy of science in a European perspective (pp. 411–424). Berlin: Springer.

Rugh, S. E., & Zinkernagel, H. (2014). A critical study of time in the universe and the multiverse. In preparation.

Sudarsky, D. (2011). Shortcomings in the understanding of why cosmological perturbations look classical. *International Journal of Modern Physics D, 20*, 509–552.

Tegmark, M. (2004). Parallel universes. In J. D. Barrow, P. C. W. Davies, & C. L. Harper (Eds.), *Science and ultimate reality* (pp. 459–491). Cambridge: Cambridge University Press.

Tegmark, M. (2005). What does inflation really predict? *Journal of Cosmology and Astroparticle Physics, 4*(1), 1–75.

Vanchurin, V., Vilenkin, A., & Winitzki, S. (2000). Predictability crisis in inflationary cosmology and its resolution. *Physical Review D, 61*, 1–17.

Vilenkin, A. (2004). Eternal inflation and chaotic terminology. Preprint http://arxiv.org/abs/gr-qc/0409055v1.

Zinkernagel, H. (2008). Did time have a beginning? *International Studies in the Philosophy of Science, 22*(3), 237–258.

Zinkernagel, H. (2011). Some trends in the philosophy of physics. *Theoria, 26*(2), 215–241.

A New View of “Fundamentality” for Time Asymmetries in Modern Physics

Daniel Wohlfarth

Abstract The goal of this article is to show that a new approach for understanding the “fundamentality” of time-asymmetries provides a possible solution to the puzzle of the arrow of radiation. This understanding is not based on the property of time-reversal invariance of fundamental laws but on the structure of the solution space of fundamental law–like equations. This new understanding of “fundamentality” implies that a fundamental time-asymmetry is a generic property of the set of possible solutions to the basic dynamic equations in classical cosmology. Moreover, I show that the arrow of radiation can be understood as a necessarily occurring by-product of the cosmological time-asymmetry, which must occur in spacetimes similar to ours.

1 A New Understanding of “Fundamentality”

The first step in this investigation is well known. It arises from the distinction between the specific properties of time-reversal invariance and time-symmetry in general. In this paper, “time-reversal invariance” is used only to describe a property of law-like dynamic equations (LLDE’s), and “time-(a)symmetry” is understood as a property of solutions to the LLDE’s. A solution f exhibits time-symmetry iff: there is at least one time point t_0 such that $f(t_0 + t) = f(t_0 - t)$ for all t. A dynamic equation $D(t)$ is time-reversal invariant iff: $D(t) = D(-t)$.

D. Wohlfarth (✉)
Institute of Philosophy, University of Bonn, Bonn, Germany

Faculty of Philosophy, University of Cambridge, Cambridge, UK
e-mail: wohlfarth@3points.de

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_23,

According to this distinction, four combinations are possible:

(a) Time-reversal invariant LLDE and only time-symmetric solutions,
(b) Time-reversal invariant LLDE and some time-asymmetric solutions,
(c) No time-reversal invariant LLDE and only time-symmetric solutions, and
(d) No time-reversal invariant LLDE and some time-asymmetric solutions.

(a) is of lesser interest, because this study seeks a definition of fundamental time-*asymmetries*.

(b) A prominent example for combination (b) is given by Maxwell's equations, which are time-reversal-invariant laws (TRIL's) describable by time-reversal invariant LLDE's. Moreover, some particular solutions of the TRIL are time-asymmetric (Jackson 1999). This is interesting because it shows that time-reversal invariant LLDE's can have time-asymmetric solutions. Traditionally, such asymmetries are not understood as fundamental time-asymmetries because they occur only in certain special models (solutions) of the LLDE and their occurrence is conditioned by boundary conditions of some kind.

The applicability of combination (c) or (d) in fundamental physics is at least problematic. It seems that non-TRIL's cannot be found within the laws of fundamental physics in the standard interpretation. Hence those combinations seem applicable only in some special formulation of quantum laws, for example in some formulations of the rigged Hilbert space approach (see, for example, Bohm et al. (1999); Bishop (2004), Castagnino et al. (2005) or Castagnino et al. (2006)). However, it will be shown below that combination (c) or (d) need not be used to understand the time-asymmetries in a fundamental manner.

In the following, I shall show that combination (b) indicates another plausible way to define '*fundamentality*' for time-asymmetries, based only on the structure of the solution space of LLDE's.

Definition (I) *Suppose L is a fundamental LLDE and S(L) is the associated solution space. I will call a time-asymmetry "fundamental" iff*

(i) *The set of all time-symmetric solutions SS(L) has significantly less elements than the set of time-asymmetric solutions SA(L). Hence, the occurring of a time-asymmetric solution would be 'typical'.*

Now, this condition depends on the understanding of '*significantly less*'. In this paper this property is seen to be fulfilled if the dimension of SS(L) is lower than that of S(L). In that case the situation is analogue to that of a plane (two-dimensional) in a three-dimensional space. The plane (SS(L)) as well as the whole space (S(L)) can contain uncountable many points (particular solutions to L). Nevertheless, the set SA(L), given by S(L) without SS(L) (SA(L) = S(L)\SS(L)), would include 'much more' solutions than SS(L). In mathematical terminology the situation would be that 'almost all' solutions to L are time-asymmetric and 'almost none' solution would be time-symmetric. The set SS(L) would be a subset of measure zero (according to an ordinary measure).

(ii) *For time-asymmetric solutions* $f(t) \in S(L)$*, the solution* $f(-t) \in S(L)$ *refers to the same physical world as does* $f(t)$*.*

Regarding condition (ii) it seems necessary to demonstrate the possibility that two distinct solutions can describe the same physical world. This possibility shall be demonstrated according to a cosmological example: Let’s assume that $f(t)$ and $f(-t)$ are solutions to Einstein-equation. In that case the following Leibniz-argument can be made:

(a) $f(t)$ does not include *intrinsic* properties that are not also included in the same way in $f(-t)$, because there are only mirrored objects (spacetimes).
(b) Both solutions are global solutions that describe spacetime as a whole, which means that there is no time parameter (or other physical parameter) outside of the objects $f(t)$ or $f(-t)$. Thus, there is no *relation* to an outstanding circumstance. Hence, it follows:
(c) Two time-mirrored spacetimes $f(t)$ and $f(-t)$ differ neither in intrinsic properties nor in any external relation. Thus, they describe the *same* physical world. The sign of t refers only to a formal non-physical and absolute background coordinate system, therefore it has no physical relevance. So condition (ii) can, in principle, be fulfilled.

Moreover, condition (ii) is important because, if $f(t) \in S(L)$ and $f(-t) \in S(L)$ describe physically different worlds, we would have to explain why only one direction (+ or − sign) occurs in nature. Only this would explain actual time-asymmetries. However, if condition (ii) holds, there is no need for such additional considerations. In order to stress this important point: condition (i) ensures that almost all solutions of an LLDE are intrinsically time-asymmetric. But this does not lead necessarily to the occurrence of time-asymmetries in the described processes or models ($f(t)$ *and also* $f(-t)$ could occur and restore the time-symmetry). However, condition (ii) ensures that the intrinsic time-asymmetries of almost all *f’s* ensures the occurring of time-asymmetric processes or models (because $f(t)$ and $f(-t)$ are *physically identical*). But the question arises whether it is possible to find a LLDE with a set of solutions which fulfilled definition I.

One example of such an equation can be found in classical cosmology. Castagnino et al. (2003a, b) have shown that Einstein’s equations produce such a situation when we make some additional assumptions. This will be reconsidered in the following section.

2 A New Suggestion

2.1 Conditions and Time-Asymmetric Spacetimes

Two crucial conditions on the set of considered spacetimes (solutions) will be used to show that, in the set of spacetimes which satisfies these conditions, a fundamental time-asymmetry is embedded in the solution space of the LLDE in classical cosmology. The conditions are as follows:

(a) In the set of considered spacetimes, cosmic-time is definable.
(b) In the set of considered spacetimes, the intrinsic dynamics is described by more variables than the scale factor alone.

I argue that the assumption of time orientability of spacetime, which is implied in condition (a), as well as (a) itself is acceptable, even if we try to define *fundamental* time-asymmetries. In general relativity (GR), the only time coordinates that appear at a fundamental level of description are the proper-times of different elementary physical systems. The assumption that is needed (if even the directions of proper-times on different world lines should be connected) is the time orientability of spacetime. Thus, if we try to deny this assumption to achieve greater generality, we cannot discuss time-asymmetries in general but only according to a single world line.

Note also that, in addition to the assumption of time orientability, we must assume that cosmic-time is definable. This is because we would otherwise have no time parameter that allows the conception of time-asymmetries *and* that is valid for more than just one world line. Thus, it seems acceptable to assume the definability of cosmic-time, which implies the time orientability of spacetime.

Next, I argue that condition (b) is also acceptable. A necessary dynamic variable for the *dynamic* description of spacetimes is the scale factor. However, if we are interested in a *physical* universe that includes matter and energy, a universe whose dynamics are described *only* by the scale factor appears to be uninteresting. The reason for this is that the dynamics of the matter and energy content of a spacetime are not describable by the dynamics of the scale factor *alone*. A dynamic universe that has *only* the scale factor as a fundamental property cannot include matter or energy as a *dynamic* property. Thus, it seems physically required to accept condition (b).

Now, to define a fundamental time-asymmetry in classical cosmology, I shall first consider the types of spacetimes that are *time-symmetric* regarding cosmic-time. In a second step, I will show that such spacetimes belong to a subset of solutions with dimension lower than the entire solutions space and that satisfies condition (i). This analysis partly follows the physical analysis of Castagnino et al. (2003a). The relevance of this cosmic-time-asymmetry for local physical processes will be shown in Sect. 3, where the arrow of radiation is shown to be understandable as a local by-product of the cosmic-time-asymmetry.

To start with, most open spacetimes are time-asymmetric. This conclusion follows from the fact that we can define a cosmic-time-asymmetry in open spacetimes according to the asymmetric behavior of the scale factor. Therefore, I do not consider open spacetimes here, even if they appear to be the right description for our particular universe. Because the crucial question is not if our particular universe includes a time-asymmetry but if the concept of a fundamental time-asymmetry is applicable to classical cosmology. In the context of classical cosmology, open spacetimes are time-asymmetric, and we seek the origin of time-symmetric spacetimes in classical cosmology.[1] Thus, to examine the origin of time-symmetry, there is no need to consider open spacetimes. According to the singularity theorems (Penrose 1979; Hawking and Ellis 1973), closed spacetimes have only one maximum in the scale factor.

For simplicity, consider the simplest case where the dynamics of all considered spacetimes are described only by scale factors $a(t)$ and scalar matter-fields $\phi(t)$. I shall argue that, in such a simple example, we can understand the origin of a fundamental time-asymmetry and we can show that this time-asymmetry is not provided by the simplifications in the example.

In Hamilton mechanics, dynamic equations depending on dynamic variables and their first derivatives in t. Thus, in this example, we have four arguments in each ‘Spacetime-Hamiltonian’: $H\left(a(t), \frac{da}{dt}, \phi(t), \frac{d\phi}{dt}\right) = 0$. Now, analytic mechanics allows us to describe one of these variables as a function of the others, and the choice of which variable depends on the others is just a matter of the description. Thus, for simplicity, I choose $a(t) = f\left(\frac{da}{dt}, \phi(t), \frac{d\phi}{dt}\right)$, where $\frac{da}{dt}, \phi(t), \frac{d\phi}{dt}$ 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, we know that $a(t)$ has just one maximum. Next we can choose the origin of cosmic-time. For simplicity, let $a(0)$ be the maximum value of the scale factor. Thus, as a function of time, $a(t)$ is symmetric with respect to the axis a at the point $t = 0$. Therefore, $\frac{da}{dt}$ is symmetric with respect to the point $\left(t = 0; \frac{da}{dt} = 0\right)$. However, for a time-symmetric spacetime, the behavior of $\phi(t)$ and $\frac{d\phi}{dt}$ *together* with $\frac{da}{dt}$ must also be symmetric. Thus, in this example, we have only two possibilities for the behavior of $\phi(t)$ and $\frac{d\phi}{dt}$ at the cosmic-time point $t = 0$, which makes the entire spacetime time-symmetric. Those possibilities are given by the triplet $\left\{\left.\frac{da}{dt}\right|_{t=0} = 0, \phi\,(t = 0)\,, \left.\frac{d\phi}{dt}\right|_{t=0} = 0\right\}$, which is a symmetric solution of $\phi(t)$ with respect to the ϕ axis at the point $t = 0$, and the triplet $\left\{\left.\frac{da}{dt}\right|_{t=0} = 0, \phi\,(t = 0) = 0, \left.\frac{d\phi}{dt}\right|_{t=0}\right\}$, which is a symmetric solution of ϕ with respect to the point $(t = 0; \phi(t = 0) = 0)$.

Hence, all symmetric solutions can be constructed using these triplets:

$$span\left\{\begin{pmatrix} \left.\dfrac{da}{dt}\right|_{t=0} = 0 \\ \phi \\ \left.\dfrac{d\phi}{dt}\right|_{t=0} = 0 \end{pmatrix}, \begin{pmatrix} \left.\dfrac{da}{dt}\right|_{t=0} = 0 \\ \phi\,(t = 0) = 0 \\ \left.\dfrac{d\phi}{dt}\right| \end{pmatrix}\right\}.$$

[1] A static universe is not considered because it requires fine tuning of the cosmological constant and the energy and matter distribution of the universe.

The complete space for solutions is instead:

$$span\left\{\begin{pmatrix}\frac{da}{dt}\\0\\0\end{pmatrix},\begin{pmatrix}0\\\phi\\0\end{pmatrix},\begin{pmatrix}0\\0\\\frac{d\phi}{dt}\end{pmatrix}\right\}.$$

Thus, the time-symmetric behavior of a spacetime is given only in a subspace that has a lower dimension than the entire solutions space *even* if we consider only closed spacetimes.

Therefore, assuming that cosmic-time can be defined and that more variables than the scale factor describe the dynamics of the considered spacetimes, we see that time-asymmetry in terms of cosmic-time is a generic property of the simplified example. But this also holds if we depart from the simplifications and add more dynamic variables, because the calculation in those cases are analogues, and the entire space of solutions always has a dimension higher than that of the subspace of time-symmetric solutions. Thus, condition (i) from definition I describes a generic property of the solution set (restricted by the mentioned condition) of the fundamental LLDE in classical cosmology.

Moreover, it is noticeable that the proposed understanding of cosmic-time-asymmetries (independent of a fundamental understanding) has a crucial advantage (see Castagnino et al. 2003a, b). This is that the cosmic-time-asymmetries are independent of thermodynamic considerations:

> Traditional discussions about the arrow of time in general involve the concept of entropy. In the cosmological context, the direction past-to-future is usually related to the direction of the gradient of the entropy function of the universe. But the definition of the entropy of the universe is a very controversial matter. Moreover, thermodynamics is a phenomenological theory. Geometrical properties of spacetime provide a more fundamental and less controversial way of defining an arrow of time for the universe as a whole. (Castagnino et al. 2003b, p. 1)

Regarding this view, the suggested understanding of fundamentality can play an additional and fruitful role. In the original view, Castagninio et al. try to argue that the geometrical structures of spacetime should be understood as a more basic property than thermodynamic properties. Even if I am very attracted by that view it seems prima facie questionable. Given that empirical-equivalent reformulations of GR (EEGR's) are possible in which the geometrical structure changed (combined with different dynamic laws), the ontic status of spacetime geometry itself is unclear. Moreover, attempts to provide certain entropic properties from specific interpretations of quantum mechanics, combined with a non-ontic understanding of spacetime geometry, can provide the opposite view, which is that entropic properties could be seen as more 'basic' than the structure of spacetime geometry.

However, the global and non-entropic considerations from Castagnino et al. (2003a, b) combined with the suggested understanding of fundamentality avoid this difficulty. In the suggested view the time-asymmetric structure of the solution set is independent from an ontic interpretation of spacetime geometry. Every EEGR with different geometries leads to an equivalent structure of the solution set (because there are empirically equivalent to GR). Hence, the fundamentality of the time-asymmetry becomes disentangled from the ontic status of spacetime geometry itself. Thus, the suggested understanding of fundamentality makes the intrinsic robustness of the cosmological account from Castagnino et al. explicit.

Back to the main subject of this paper; it was shown that condition (i) from definition I if fulfilled for the crucial solution set in classical cosmology. Thus, we have to examine whether condition (ii) is also satisfied.

2.2 Solution Space

For mathematical reasons, the solution space is built from time-mirrored pair functions $f(t)$ and $f(-t)$. This means that each time-asymmetric solution $f(t)$ has a pair function $f(-t)$ that is also a solution to the LLDE. Almost all of them are intrinsically time-asymmetric (condition (i)), but the directions of the asymmetries seem to have been mirrored. Here the phrase ‘mirrored’ refers only to the formal fact that the sign of t switches, so $f(\pm\, t)$ is a ‘t-mirrored’ solution-pair to the LLDE. Hence, definition I is not shown to be fulfilled so far. But, according to the Leibniz argument in Sect. 1, it is shown that the two time-mirrored spacetimes $f(t)$ and $f(-t)$ differ neither in intrinsic properties nor in any external relation. Hence, they describe the *same* physical world.

Thus, the solution space is *not* built from physically different time-mirrored pairs. Note, however, that the proposed approach works only for *global* solutions of *global* LLDE’s. In the context of, for example, particle physics, where the CPT theorem has been developed, we normally discuss the transformation of particle properties or systems constructed from particles. In this case the t-mirroring is physically important, as the adequacy of the CPT theorem shows (Gross 2004).

Nevertheless, the conclusion is that the solution space of the LLDE’s which describes possible spacetimes as a *whole* does not consist of distinct *physical* pairs $f(t)$ and $f(-t)$. Thus, we have found a *fundamental* [definition (I)] time-asymmetry in cosmology with explications in almost all models which are fulfilling conditions (a) and (b) from Sect. 2.1. I shall thus argue that the prominent arrow of radiation can really be understood as a necessarily occurring local consequence of the fundamental cosmic-time-asymmetry.

3 The Arrow of Radiation

I characterize the arrow of radiation as described in standard textbooks (Jackson 1999) and research literature (Rohrlich 2005; Jauch and Rohrlich 1976) or (Frisch 2000) as the absence of fully advanced radiation. But it seems necessary to distinguish possible kinds of fully advanced radiation, which are occasionally discussed in this context:

(a) Source-free fields coming from '± infinity' can be combined in a way which makes the fully advanced description applicable to the phenomenon.
(b) Fully retarded emitters can be arranged in a special geometry such that the combined field becomes describable, in a special region, as fully advanced.
(c) An accelerated charge is supposed to radiate but the associated radiation field could be fully advanced or fully retarded. In nature only (or almost) the fully retarded solution to Maxwell's equation seem to occur.

I shall argue that only the non-occurrence of fully advanced radiation from type (c) provides a suitable characterisation of the arrow of radiation. Consider fully advanced radiation from type (a). Given the observations we detect quasi source-free radiation from the microwave background but no radiation can be observed coming from the cosmic future. This could be interpreted as a time-asymmetry or not (see for example Price 2006). Nonetheless, given the cosmological models and the observational well established assumption that our particular cosmic domain (or the observable universe) accelerates his expansion, the non-occurrence of microwave radiation from the cosmic future seems not too surprising. Even if the lack of this radiation would be interpreted as an observable time-asymmetry, it seems to be a time-asymmetry which is grounded on cosmological boundary conditions. This time-asymmetry will therefore not be understood as the arrow of radiation in this investigation. Instead, this time-asymmetry will be understood as a consequence of the accelerated expansion of our particular cosmic domain or universe (which is not *fundamentally* given). Also, in electromagnetic shielded regions, this asymmetry is non-existing. Nevertheless, fully advanced radiation from type (c) does still not occur.

Second, fully advanced radiation from type (b) is not a special phenomenon of electromagnetic waves. All types of classical waves show this type of behaviour. A special geometry of fully retarded emitters can provide a wave field that converges coherently (in a special region) and can be described as a fully advanced wave field (in that region). But the *total* wave field of the retarded emitters is not appropriately described as fully advanced. This possibility of special emitter geometries is not connected to a suitable characterisation of the arrow of radiation. The special emitting geometries are built of fully retarded emitters. Thus, the fundamental emitters are fully retarded and are merely arranged so as to produce a radiation field that can be described (in a special region) as fully advanced. Hence, the arrangement of the emitters is crucial and in nature the absence of such arrangements is well understood (even if not in fundamental terms) by thermodynamic considerations (see for example Popper 1956 or Price 2006).

But there seems to be another possible kind of fully advanced radiation, which is not observed in nature. The absence of this radiation, type (c), seems to provide a basic time-asymmetry in classical electrodynamics (not obviously provided from thermodynamic considerations or boundaries) and hence should be used to characterise the term 'arrow of radiation'. This, I think, avoids the confusion that could occur if fully advanced radiation from type (a) or (b) is taken into account. Thus, in this investigation, the term 'arrow of radiation' will be understood as the fact that radiating accelerated charges seems not (or only rarely) to be associated with fully advanced radiation, even in electromagnetic shielded regions.

Many attempts have been made to solve the puzzle of the arrow of radiation (Price 1996; Frisch 2000, 2005, 2006; Zeh 2010), both in philosophy as well as in physics and sometimes by arguing for a different characterization of the time arrow itself (Price 1996). Accounts which favor a different characterization are not discussed in any detail here because of the given motivation for the proposed characterization. But other authors suggested that the absence of fully advanced radiation is based on a new law of classical electrodynamics (Frisch 2000) or on an assumed fundamental time-asymmetry of causation (Frisch 2005; Jackson 1999; Rohrlich 2005). Because the goal of this paper is not to describe the relative merits of the different approaches but to suggest a new understanding of the radiation arrow, I will only sketch possible critiques of these popular attempts.

For example Frisch 2000 who suggested a new law:

> The account I wish to advocate simply stipulates that, in addition to the Maxwell equation, electromagnetic fields associated with electric charges satisfy the retardation condition without offering any explanation as to why this condition should hold (Frisch 2000, p. 25).

With respect to that account, it seems that we should favor an account that can explain the absence of fully advanced radiation without proposing a new ad hoc law.

Also, it seems problematic to base the arrow of radiation on an assumed time-asymmetry of causation, as does Rohrlich 2005:

> The latter [fully advanced radiation] is excluded because it would have to come from sources in the future going in the negative-time direction and arriving at the particle on a future light cone. This violates causality (Rohrlich 2005, p. 3).

In this view, causality should provide a time-asymmetry in physical processes even if physics is unable to explain why causation should be time-asymmetric (see e.g., Price 1996). Moreover, if the time-asymmetry of causation is not provided from physics, it is not obvious how physical processes can be guided by that time-asymmetry.

Instead of going into detail in these interesting discussions, I suggest a new account that shows that the absence of fully advanced radiation (type c) can be explained as a consequence of the fundamental time-asymmetry in the solution set of possible spacetimes. In order to do so, the first issue to address is that the cosmic-time-asymmetry is an asymmetry of cosmic-time, whereas the arrow of radiation refers to proper-times. Therefore, I deduce the local asymmetry which refers to proper-times from the global asymmetry. More precisely, I show how to distinguish semi-light-cones (in one time-asymmetric spacetime) in a non-conventional way.

It is well known (Earman 1974) that we can use a non-vanishing, continuous timelike vector field on a time orientable spacetime to distinguish between semi-light-cones:

> 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).

With a non-vanishing, continuous timelike vector field, we can understand that the difference between past and future semi-light-cones is non-conventional, because this difference is given by the physical difference between the *cosmic-time directions* in almost all spacetimes. According to the cosmic-time-asymmetry the difference between the proper-time directions is also physical, because one semi-light-cone contains all the timelike vectors pointing in one of these cosmic-time directions, whilst the other semi-light-cone contains all the timelike vectors pointing to a *physically different* cosmic-time direction.

Thus, the difference between future and past semi-light-cones would be given by the direction of this timelike vector field, and this direction stays the same at each point in spacetime. However, at first glance, such a difference appears to be more technical than physical, because the vector field is, as presented, just a mathematical construction. Therefore, the next step is to identify physical candidates to play the role of the continuous, non-vanishing and timelike vector field. As we will see below, it is useful to consider the energy-momentum tensor:

$$T_{\mu\nu} = \frac{1}{8\pi}\left(R_{\mu\nu}(g) - \frac{1}{2}g_{\mu\nu}R(g) - \Lambda g_{\mu\nu}\right). \tag{1}$$

The components of $T_{\mu\nu}$, as they appear in (Eq. 1), do not play the role of a continuous, non-vanishing timelike vector field.[2] However, we can add two conditions to $T_{\mu\nu}$ which seem to be fulfilled in our particular spacetime:

(a) $T_{\mu\nu}$ is a type-I energy-momentum tensor.[3]
(b) $T_{\mu\nu}$ satisfies the dominant energy condition $T^{00} \geq |T^{\mu\nu}|$ for any orthonormal basis.

In this case, with condition (a), we can write Eq. 1 in the form

$$T_{\mu\nu} = s_0 V^0_\mu V^0_\nu + \sum_{i=1}^{3} s_i V^i_\mu V^i_\nu, \tag{2}$$

where $\{V_\mu{}^0, V_\mu{}^i\}$ is an orthonormal tetrad and, as in the standard notation, $V_\mu{}^0$ is timelike and $V_\mu{}^i$ is spacelike with $i \in \{1,2,3\}$.

[2] $R_{\mu\nu}$ is the Ricci tensor, R the Ricci curvature, Λ the cosmological constant, and $g_{\mu\nu}$ the metrical tensor.

[3] This means describable in normal orthogonal coordinates. See Hawking and Ellis (1973) and also Eq. 2.

Additionally, (b) shows that s0 ≥ 0 and that si is given by $s0$ or $-s0$. Therefore:

$V_\mu{}^0(x)$ (where x gives the spacetime coordinates) is a continuous, non-vanishing timelike vector field. Moreover, $T^{0\mu}$ can be interpreted as the physical energy flux, described by a continuous, non-vanishing timelike vector field.[4]

But note that we cannot make any assumptions about the type of the energy-momentum tensor in general, because nothing is known about the phenomenology of other possible universes, and I will not consider speculations about quantum gravity. As a consequence, all results are restricted to spacetimes which fulfill the mentioned conditions on the energy-momentum tensor. However, the conclusions are nevertheless interesting because they help to understand the nature of the arrow of radiation, by determining the condition on which the absence of fully advanced radiation depends.

Given the interpretation of $T^{0\mu}$ as the time-asymmetric energy flux, energy flows always from the proper past to the proper future (which is fundamentally different from the proper past by the light-cone-structure). Thus, fully advanced radiation (type c) is not possible by this asymmetry in the energy flux, because this type of fully advanced radiation would imply an energy flux from the proper future to the proper past. Thus, the existence of the arrow of radiation is explained as coming from two separate parts: first, there is a fundamental time-asymmetry in the solution set of the LLDE's in classical cosmology which provides an explication in almost all cosmological models, a time-asymmetry with respect to cosmic-time. But this asymmetry of cosmic-time is insufficient to explain the existence of the radiation arrow. The second part of the explanation consists of the crucial conditions (a) and (b), which lead to local and time-directed consequences in classical electrodynamics.

Thus, a concept for the origin of the arrow of radiation results in which the arrow originates in the interaction between fundamental time-asymmetric cosmology and some conditions (fulfilled in our particular spacetime) on the energy-momentum tensor.

References

Barcel'o, C., & Visser, M. (2002). Twilight for the energy conditions? *arXiv:hep-th/0205066v2*.

Bishop, R. C. (2004). Arrow of time in rigged Hilbert space quantum mechanics. *International Journal of Theoretical Physics, 43*(7/8), 1675–1687.

Bohm, A., Gadella, M., & Wickramasekara, S. (1999). Some little things about rigged Hilbert spaces and quantum mechanics and all that. In I. Antoniou & G. Lumer (Eds.), *Generalized functions, operator theory, and dynamical systems* (pp. 202–250). Boca Raton: Chapman & Hall/CRC.

[4] This interpretation appears to be canonical in the context of GR, but there are exceptions. Nevertheless, the exceptions come only into play by considering quantum cosmology and quantum field theory. Critical points are, for example, the Casimir effect, the squeezed vacuum, or Hawking evaporation (see, e.g., Visser 1996 or Barcel'o and Visser 2002).

Castagnino, M., Lombardi, O., & Lara, L. (2003a). The cosmological origin of time asymmetries. *Classical and Quantum Gravity, 20*, 369–391.

Castagnino, M., Lombardi, O., & Lara, L. (2003b). The global arrow of time as a geometrical property of the universe. *Foundations of Physics, 33*, 877–912.

Castagnino, M., Gadella, M., & Lombardi, O. (2005). Time's arrow and irreversibility in time-asymmetric quantum mechanics. *International Studies in the Philosophy of Science, 19*, 223–243.

Castagnino, M., Gadella, M., & Lombardi, O. (2006). Time-reversal, irreversibility and arrow of time in quantum mechanics. *Foundations of Physics, 36*, 407–426.

Earman, J. (1974). An attempt to add a little direction "The problem of the direction of time". *Philosophy of Science, 41*, 15–47.

Frisch, M. (2000). (Dis) solving the puzzle of the arrow of radiation. *The British Journal for the Philosophy of Science, 51*, 381–410.

Frisch, M. (2005). Counterfactuals and the past hypothesis. *Philosophy of Science, 72*, 739–750.

Frisch, M. (2006). A tale of two arrows. *Studies in the History and Philosophy of Modern Physics, 37*, 542–558.

Gross, F. (2004). *Relativistic quantum mechanics and field theory*. New York: Wiley.

Hawking, S. W., & Ellis, G. F. R. (1973). *The large scale structure of space-time*. Cambridge: Cambridge University Press.

Jackson, J. D. (1999). *Classical electrodynamics*. Berlin: Walter de Gruyter.

Jauch, J. M., & Rohrlich, F. (1976). *Theory of photons and electrons* (2nd ed.). New York: Springer-Verlag.

Landau, L., & Lifshitz, E. (1970). *Theorie des champs*. Moscow: Mir.

Penrose, R. (1979). Singularities and time asymmetry. In S. W. Hawking & W. Israel (Eds.), *General relativity: An Einstein centenary survey* (pp. 581–638). Cambridge: Cambridge University Press.

Popper, K. (1956). The arrow of time. *Nature, 177*, 538.

Price, H. (1996). *Time's arrow and Archemede's point*. New York, Oxford: Oxford University Press.

Price, H. (2006). Recent work on the arrow of radiation. *Studies in the History and Philosophy of Modern Physics, 37*, 498–527.

Rohrlich, F. (2005). Time reversal invariance and the arrow of time in classical electrodynamics. *Physical Review E, 72*, 057601.

Visser, M. (1996). *Lorentzian wormholes*. Berlin: Springer.

Zeh, H. D. (2010). *The physical basis of the direction of time*. Berlin: Spinger.

Part VII
Philosophy of the Physical Sciences: From Physics to Metaphysics

How to Combine and Not to Combine Physics and Metaphysics

Mauro Dorato

Abstract In this paper I will argue that if physics is to become a coherent metaphysics of nature it needs an “interpretation”. As I understand it, an interpretation of a physical theory amounts to offering (1) a precise formulation of its ontological claims and (2) a clear account of how such claims are related to the world of our experience. Notably, metaphysics enters importantly in both tasks: in (1), because interpreting our best physical theories requires going beyond a merely instrumentalist view of science; in (2), because a philosophical elaboration of the theories of the world that are implicit in our experience is one of the tasks of analytic metaphysics, and bridging possible explanatory gaps or even conflicts between the physical image and the manifest image of the world (Sellars 1962), which is a typical philosophical task that involves both science and metaphysics.

In order to defend this claim, in Sect. 1 I attack a widespread position about the relationship between metaphysics and science, namely the view that metaphysics is to be regarded as the a priori study of a space of possibilities, and science may enter in choosing among alternative accounts of such a space. In Sect. 2 I briefly criticize contemporary forms of physicalist chauvinism, which are mirrored by dual attitudes on the part of the metaphysicians who ignore science, and then present my own view of the relationship between metaphysics and physics.

M. Dorato (✉)
Department of Philosophy, Università degli Studi Roma 3, Viale Ostiense 234,
00144 Rome, Italy
e-mail: dorato@uniroma3.it

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_24,

1 Metaphysics as the Study of the Possible

An influential position in contemporary debate is that metaphysics is concerned with the study of a space of possibilities (Lowe 1998, 2011). The idea is that metaphysics studies the world not as it *actually* is (which is the task of science), but as *it might be*.

Four remarks are in place in order to further clarify this position.

1) First, as French and McKenzie stress (2012), this view, with its stress on "possibility", has received a lot of momentum, as much recent metaphysics, from Kripke's revival of modal logic, with its subsequent emphasis on modal metaphysics.
2) Second, this study of a space of possibility is typically presented as something to be conducted purely a priori, so that the distinction between metaphysics and physics ought to be grounded in the distinction between a priori and a posteriori methods of gaining knowledge. This fact is supposed to warrant the autonomy of metaphysics from science.
3) Thirdly, despite such an autonomy of methods, science is regarded as relevant to metaphysical inquiries, since the former might intervene in evaluating competing accounts of "the way the world might be" suggested by the latter.[1]

   If one looks at disputes such as presentism versus eternalism, perdurantism versus endurantism, haecceitistic versus reductionist view of individuality, one can easily see that relativity and quantum mechanics have often been brought to bear in order to decide between these competing views. By guaranteeing at the same time *some* degree of autonomy but also some form interaction with science, isn't this conception of the relationship between metaphysics and physics providing us with the best of all possible worlds?
4) Fourthly, and crucially in my view, the viability of this approach to metaphysics is predicated on the existence of a domain of modality which is intermediate between merely *logical* possibility (absence of contradiction) and *nomological* possibility, namely an intermediate domain of *metaphysical* possibility.

In order to show that this conception of metaphysics regarded as the study of a space of possibility suffers from many objections, I will concentrate my attention on the third and the fourth point, while commenting briefly on each of the preceding two points.

1. Historically, one might think that when modern natural philosophers rejected modality (the "essences") together with Aristotelianism, they threw the baby out with the bath water. And clearly, in the last 50 years, the availability of a rigorous semantics for modal statements has opened new pathways to metaphysical speculation. David Lewis' work, in particular, has been very influential in shaping much contemporary discussion in metaphysics, a discussion not always influenced by what was going on at the same time in the sciences. No neutral judgement can be passed on this modal trend of contemporary analytic

[1] This point has been recently stressed by Morganti (2013).

metaphysics, since a Carnapian and a Quinean will judge it as a deplorable tendency of contemporary philosophy, while their opponents will welcome it. Certainly, it is curious to see that contemporary metaphysicians re-discovered a lot of Latin-derived categories of medieval philosophy, not only essentialism, but also haecceitism, quidditism, potentiae (powers) and the like, a fact that, without further analysis, need not entail that we are returning to scholastic philosophy in its pejorative sense. A most liberal attitude might end up granting metaphysicians all the freedom they need to pursue metaphysical research programs that today appear totally disconnected from science. We cannot exclude that eventually the toolbox provided by these inquiries may prove useful for the philosopher of physics (French and McKenzie 2012, p. 43) and therefore also for physics itself, at least to the extent that, by paraphrasing Chang, the "philosophy (and history) of physics is the continuation of physics with other means" (Chang 2004, p. 235). It would be interesting to provide some further historical evidence for this "applicability argument", in the same sense in which the history of science has uncontroversially shown that pure mathematics has proved immensely useful for the empirical progress of physics. But this is something that cannot be pursued here.

2. On the contrary, the history of the twentieth century philosophy of science has, I take it, provided good and abundant evidence that metaphysics cannot easily and precisely be *demarcated* from science. Why should the a priori/a posteriori criterion succeed where Wittgenstein, Popper and their followers failed? The a priori character of metaphysics is of course an important trait of its method, at least after the semantic turn (Coffa 1993), and clearly depends on its tendency to analyse the *meaning* of key notions (Boghossian and Peacocke 2000).

However, this aprioristic trait is not sufficient to distinguish it from science and grant full autonomy from it. Mathematics is certainly a science but mainly justified a priori, and many empirical sciences rely on mathematical models of phenomena which exist, if they do, in an abstract dimension that might nevertheless be part of the ontology of science (Psillos 2011). Furthermore, there are significant instances of conceptual, a priori analysis also within the empirical sciences: Einstein's 1905 analysis of the meaning of simultaneity is only one of the most famous examples. On the other hand, some contemporary philosophy has called upon the experimental method ("experimental philosophy") in order to evaluate the credibility and strength of some philosophical intuitions. Of course, this does not mean that metaphysics and science cannot be distinguished, at least *prima facie*, by looking at their different epistemological methods. But it does mean that no clear-cut demarcation can be drawn between science and metaphysics simply by looking at the a priori/a posteriori distinction. The next point casts additional doubts on the possibility of distinguishing physics from metaphysics simply in terms of the a priori/a posteriori divide.

3. The idea that science becomes relevant to metaphysics when it is invoked to choose between alternative conceptual characterizations of possibility spaces seems to render metaphysics open to empirical refutation.

However, notice the following dilemma, which is relevant, at least in part, also to the preceding point (2). *Either* science is incapable of providing some evidence for a metaphysical theory, and metaphysics is therefore epistemically wholly autonomous from science, or metaphysics cannot be wholly divorced from the a posteriori side of the process of gaining scientific knowledge, since the ultimate justification for a metaphysical claim comes from *empirical* considerations. By choosing the first horn, we gain the autonomy of metaphysics at the expense of its relevance for science. By taking the second horn, metaphysics cannot be sharply distinguished from science just by looking at its epistemology, so that the distinction between a priori metaphysics and a posteriori science is not waterproof. The cost of choosing the first horn seems *prima facie* too high, especially for those who, like myself, want to advocate a form of interaction between science and metaphysics. The second alternative, however, to be further explored in what follows, will also prove unsatisfactory, at least in terms of its capacity to cross-fertilize physics with metaphysics. Since both horns are unsatisfactory, I will conclude that the whole conception of metaphysics regarded as the analysis of a space of possibility should be jettisoned.

First of all, let us remark that the second horn is compatible with the claim that the process of construction of a metaphysical theory is wholly a priori, so that it is only its validation or final justification that is a posteriori. This validation requires at least a *compatibility test*: if a metaphysical theory is in conflict with a well-confirmed physical theory, the former ought to be abandoned. Compatibility tests are frequently invoked to introduce a wished-for interaction with physical theories: metaphysical theories are in the same relation with respect to science, as scientific theories are with respect to experiments.

Unfortunately, this way of construing the relationship between physics and metaphysics is rather weak, because it is open to the following two objections: (1) it does not create a fertile interaction between physics and metaphysics and (2) it leads to the claim that metaphysical theories are underdetermined by science. By discussing a couple of case studies taken from the philosophy of time, I will now illustrate both these objections, that I label "sterility" and "underdetermination".

3.1 As to sterility, consider the dispute between *presentism* (advocating that all and only present events exist) and *eternalism* (according to which past present and future events are ontologically on a par), or that between *perdurantism* (entities have temporal parts) and *endurantism* (entities are wholly present at each instant of their existence). As an argument in favour of the sterility objection, note that these two metaphysical debates, *even granting that they are genuine*,[2] are somehow completely *external to physics*. They are external because current physicists do not care at all about the question whether the future is real or not, or whether entities endure or perdure, even though it can

[2] For reasons against the genuineness of the presentism/eternalism debate, see Dolev (2006), Dorato (2006a), Savitt (2006). Against the genuineness of the endurantism/perdurantism debate, see Dorato (2012).

be admitted that eternalism and perdurantism are closer to the requirements of Minkowski spacetime. Unlike, say, the question of the origin of the arrow of time, these two debates in the philosophy of time are not open, debated problems of contemporary physics: this, clearly, is *not* to say that these two metaphysical problems of time are philosophically uninteresting, but it is to say that whoever is concerned with creating a fruitful interaction between physics and metaphysics will remain disappointed.

3.2 The underdetermination of metaphysics by physics in our two case studies originates because if, for *metaphysical* reasons, one is willing to add *an empirically inaccessible* inertial frame to Minkowski spacetime, strictly speaking one is not contradicting at all special relativity as a *physical* theory. It is only for methodological and, in the last analysis, philosophical reasons that we prefer special relativity without a privileged, but empirically inaccessible, frame: an inaccessible preferred frame is a difference that does not make any empirical difference. But this philosophical reason *can* be overcome in the name of other *philosophical* reasons.[3] Analogously, if someone were interested in claiming that entities endure, that is, they persist in time by being wholly present at each moment in which they exist, she will be ready to pay the price of introducing a privileged frame, playing somehow the role of an empirically inaccessible present, providing the three-dimensionalism that is needed for the corresponding metaphysical view. And in fact, this is what frequently happens among the defenders of presentism (Craig 2001) and perdurantism.

Sterility and underdetermination are two reasons why using physics as an experimental test for competing metaphysics theories is insufficient to create a fruitful interaction between the two disciplines. This verdict depends of course on the particular metaphysical theory we are discussing, but if sterility and underdetermination were correct for *many* metaphysical debates, we would be pushed back to the first horn of the dilemma: physics cannot provide decisive evidence for a metaphysical theory.

4. The fourth problem involves the availability of a notion of metaphysical modality (necessity, possibility) that is intermediate between mere logical possibility and nomological possibility. It should be evident why the notion of merely logical possibility is too weak to produce metaphysically interesting claims. Absence of contradiction is necessary and sufficient to build a logically possible world, but it is insufficient for a serious inquiry on metaphysical possibility. It is certainly non-contradictory to imagine or conceive an individual that is cell by cell identical to me but deprived of mental states, but this argument based on conceivability is not illuminating on the nature of mental states or on body-mind relationship, especially if it proved to be *nomically impossible* to have mental states without a *physical* realization of some sort.

[3] The case of Bohmian mechanics or of some versions of the GRW dynamical reduction theory is different, since these alternative theories need an additional frame for more "physical" reasons.

Even if what has just been said were unconvincing, the problem of defining a metaphysical possibility that is independent on, or at least significantly autonomous from, nomological possibility, which is the object of science, should be solved *before* assigning metaphysics the task of "opening possibility spaces". In fact, another way of posing the question of the relationship between science and metaphysics is to ask oneself whether and to what extent *metaphysical* possibilities or necessities are independent of the corresponding *nomic* modalities that are the object of scientific investigation.

It could be noted that even if metaphysical necessities were supervenient on the nomological necessities fixed by the laws of nature, there might be problems that are scientifically open or even scientifically unsolvable, problems that are nevertheless important and that can be tackled only by philosophy. This must be granted. But if we (currently or in principle) ignore what all the laws are if there are laws regulating phenomena that are (currently or in principle) underdetermined by our scientific knowledge, it then seems to follow that the individuation of metaphysical possibilities becomes epistemically dependent on the recognition of *merely logical possibilities*, with the limitations mentioned above. These limitations become particularly evident if, following Chalmers (2002, p. 13), metaphysical possibilities and necessities are regarded as corresponding, more or less, to *ideal conceivability*: in trying to discover logical possibilities, it is conceivability that is typically advocated. It therefore seems reasonable to conclude that it is highly dubious that there is an intermediate modality between logical and physical possibility.

If *both* metaphysics and physics are attempts at describing the general structure of reality, and there cannot be a "double truth" about reality, and nor can there be a "double method" – one empirical and one a priori – in order to find out the way the world is, we should look for better meta-metaphysical theories, and abandon the view that metaphysics is the study of logical possibility.

2 Metaphysics, Physics and the Nature of Interpretation

One radical solution of the problem of the relationship between physics and metaphysics is denying the necessity of one of the two *relata*.

Physicalistic Chauvinism, for one, is the claim that physics, *being itself a metaphysics of natur*e, does not need any contribution from an "external" philosophy. The idea here has been well expressed by DiSalle in an historical context (he does not endorse it in the way I present it): "physics … *is* the metaphysics of nature. The metaphysical concepts that we find in physics – body, force, motion, space, time, become to us intelligible precisely, and only, as they are constructed by physics itself; physics provide us with the only intelligible notions we have on this matter" (DiSalle 2006, p. 60). This form of chauvinism is justified by the historical facts that (i) physics keeps on appropriating key concepts that were previously part of metaphysics, and that (ii) concepts of the manifest image (particle, wave etc.) are often incapable of applying to physical areas of investigations that, like quantum mechanics, are very remote from the macroscopic world of our experience.

However, this attitude, mirrored by the dual, “proud” ignorance of science that some metaphysician profess and that has been amply illustrated by Ladyman and Ross (2007), seems a bit too autarchic: *since*, as I will argue in the remainder of this essay, “philosophers’ metaphysical theories are an elaboration of the manifest image, physics, if it is not to be reduced to a mere cookbook for predictions, has the task of connecting with such an image, since all its evidential force, after all, comes from experience. While a subordination of science to metaphysics is today unthinkable, once again the question becomes whether, before yielding to physical chauvinism or scientism, we can find a more fruitful way to have physics interact with metaphysics.

One very effective way to achieve this aim is via the two-layered task of *interpreting* physical theories, which means: (1) coming up with a precise and exact *ontology* to associate to the language and formulas of physical theories and (2) relating such an ontology to the world of our experience. It then seems that project (1) necessarily involves a metaphysical task, namely finding out how the world can be like *if* our physical theories are (at least approximately) true.

Project (1) has been variously defended (van Fraassen 1980; Giere 1988; Lange 2002) and does not require a special argument here. However, in order to realize the importance of (1), its relation to (2) is essential, and *this* aspect has *not* been sufficiently stressed. In what follows I present one remark (R), an historical reflection (HR) showing the centrality of the relationship between (1) and (2) and two examples (E1 and E2) supporting the view that connecting (1) with (2) offers key suggestions also for relating metaphysics and physics.

R) The question of the often conflicting relationship of the ontology of physics with that of our experience arises only if the former is taken seriously. It is only if the table is *really* made of atoms and light of electromagnetic radiation that the question of the relationship of the “empty physical table” with the hard and coloured table of our experience becomes serious (Eddington 1928, p. ix). This is no argument for scientific realism, of course. All I want to claim is that since physics *could* be interpreted realistically, it ought to explain away any source of conflict with the manifest image, since also instrumentalists recognize that all its evidential force comes from observations belonging to the world of our experience. In the hypothesis that there is only one table and two descriptions of it, the attempt at linking together in a harmonious whole the scientific and the manifest descriptions seems worthy of the “synoptic” work typical of philosophers. Such an attempt calls into play the cognitive and neural sciences, evolutionary psychology, the philosophy of language and the history of science, and not just *aprioristic* conceptual analysis, even though such an analysis is indispensible in order to clarify the implications of our manifest image: this clarification is achieved via explications of key concepts like object, event, property, causation and the like.

HR) An analysis of the conflict between the ontology of physics and the manifest image has been decisive in revolutionary changes, when categories that were central in our manifest image had to be abandoned because they stood in the way of gaining a clear understanding of the physical phenomena. In particular, in various revolutionary changes of the past we have abandoned a search for *causes*. Think of: (i) pre-Galileian attempts at causally *explaining* what we now call inertial motion by invoking motive powers, suggested by the role that friction plays in our experience; of (ii) pre-Einstenian attempts at explaining the invariance of the speed of light by presupposing a length contraction due to intermolecular *forces*; or (iii) of the Newtonian postulation of an unobservable gravitational force to explain free fall, which nowadays we consider to be locally equivalent to *inertial motion*. We now know that inertial motion, the invariance of the speed of light and free fall are fundamental and "natural", in the sense that they need no causal explanations whatsoever. However, experiments in naïve physics tell us that we still perceive falling objects in an Aristotelian fashion (we don't perceive their acceleration), and this explains why we naïvely tend to explain certain phenomena by presupposing the world of the manifest image (McCloskey et al. 1980). Analogously, against relativity, we spontaneously believe in a cosmic present and therefore in absolute simultaneity, and tend to presuppose the notion of cause/force in order to explain motion.

These historical examples show that a dialectic between the scientific and the manifest image was at the heart of each revolutionary theory of the past; and this fact might prove important also to understand which part of the manifest image we must give up in order to achieve cognitive progress in the future. For instance, shouldn't we abandon causal explanations of non-locality and regard quantum correlations as fundamental? The crucial novelty yielded by quantum mechanics – that we have had such a hard time to understand relative to the manifest image – is precisely *entanglement*. The quantum correlations ought just to be regarded as fundamental as inertia, the invariance of light and free fall: as such they need no causal explanation at all (Fine 1989).

In stubbornly looking for causal models for the quantum correlations, we seem to apply our manifest image to a scale that is far too remote from the environment to which we adapted. What needs an explanation is not entanglement – which in the quantum world can be taken as fundamental – but rather why we don't perceive macroscopic superpositions, and therefore entanglement, at the macrolevel. This epistemic switch requires that entanglement be regarded as the main ontological lesson of the quantum world, in such a way that it can be presupposed to explain our experience of macroscopic definiteness. If entanglement were not part of the ontology of the physical world in my sense of "interpretation", quantum mechanics would have nothing to explain *vis à vis* the macroscopic world of our senses; but this claim is regarded as false by many practitioners of the subject.

E1) Such a falsity is particularly evident in all of the interpretations of quantum mechanics, in particular in the no-collapse views related to Everett. By denying the reality of the reduction process, Everett's approach must still explain the *appearance* of such a process, and therefore must face the problem of interpretation in my sense: the metaphysical posit here consists in claiming that the universe is described by an evolution equation which is always linear, time-symmetric and deterministic.

Under this ontological presupposition, two correlated problems arise, both involving consistency with what we see, and therefore *the relationship between the relevant metaphysical posits and the manifest image*. The first problem is why, despite the lack of a genuine collapse, we never perceive macroscopic superpositions. The second problem consists in trying to explain the origin of the notion of (time-asymmetric) conditional probabilities in a deterministic time-symmetric theory; namely the *impression* that the irreversible probabilities involved by the Born-rule play a fundamental role in quantum theory.

The first problem is tackled with the theory of decoherence, which explains why local observers can never perceive interferences (from within the same "world") of Schrödinger's infamous dead cat with its living counterpart, even though all possible measurement outcomes do occur. This implies that there is an observer perceiving a live cat in a world, and the "same" observer looking at the same dead cat in worlds that don't interfere with each other. The second difficulty is attacked by invoking decision theoretic strategies of *agents* (see among others, Deutsch 1999 and Wallace 2007): again, it is the robustness of the explanatory link posited between a metaphysical interpretation of a physical theory (many worlds) and our experience that gives us the final test for the plausibility of a proposed ontological interpretation of a physical theory.

The appropriateness of this interpretation of quantum mechanics of course cannot be judged in this context. Here it has been mentioned simply in order to show how complicated the interpretation of a physical theory really is, and how promising the philosophical program sketched here really is, if one cares about having physics and metaphysics interact in a non-superficial way.

It should be added that the historical cases briefly alluded to above should not be taken to suggest that causal explanations are to be abandoned in *all* areas of physics. This would be too hasty, precisely because causation has such a central role in the manifest image. The task ahead is to understand in a more precise way the working of our brains and mind, and the way they construct the manifest image. It is in this sense that analytic metaphysics is indispensible to get a firmer grasp on the sort of assumptions that we unconsciously and pre-scientifically make about the outer and inner workings of the world.

E2) A final example that well illustrates the problems raised by Sellars' two images is offered by the following question: can the timelike-separation of events in spacetime theories be interpreted as giving rise to a tenseless form of *local, non-global becoming*? Philosophers who have recently advocated this minimalist claim (Savitt 2002; Dieks 2006a; Dorato 2006b) are well aware that the question remains whether such a *metaphysical interpretation* of relativity is capable of explaining the sense of the passage of time typical of our manifest image, which is exactly the explanatory task required by (2) above. If this task is not fulfilled, the ontological posit presupposed by (1) must be abandoned or at least corrected. Explaining why we "falsely" or "approximately" believe in a cosmic present extending across space is part of such an explanation and presumably needs some connection between the remarkable speed of light (a physical fact) and our limited capacity for discriminating two light signals as being successive in time, a psycho-physiological fact pointing to a threshold of about 30 ms. Since 300.000 km/s (the speed of light *in vacuo*) times 30 ms is 9,000 km, and since within a sphere of that radius a pair of light signals cannot be perceived as temporally successive by humans located in the centre of it, we have thereby a possible explanatory connection between the ontology of relativity – the partial timelike succession of events metaphysically interpreted as local becoming – and our experience of the world. If the connection were robust, we would have explained away our impression of a cosmic present constituted by absolutely simultaneous events (see also Butterfield 1984 and Callender 2008).

In conclusion, I hope to have shown that the question of interpretation in the sense above is in fact not external to physics, at least to the extent that in the past also physicists have asked themselves whether, for instance, the crystalline spheres, atoms or the ether really existed. In any cases, a precise ontological interpretation of a theory is needed to link the physical image with the world of our experience, an explanatory link which is not only important for the coherence of the physical image but is also one of the main tasks of philosophy. Studying this link takes us closer to Plato's ideal of the philosopher as capable of "syn-opsis", which is the act of looking "at all the ideas at once". Sellars' (1962, p. 36) appropriate metaphor is the sense of depth yielded by binocular vision, which results from fusing the vision of one eye (the manifest image's) with the different perspective produced by the other eye (the scientific image). If philosophy gives up this synoptic or "deep" vocation, then I fear that it is not worth the candle.

Acknowledgement I thank the editors for some comments that improved the manuscript. The paper has been written with the help of the grant FIRB/FIRB Project F81J12000430001.

References

Boghossian, P., & Peacocke, C. (Eds.). (2000). *New essays on the a priori*. Oxford: Oxford University Press.

Butterfield, J. (1984). Seeing the present. *Mind, 93*, 161–176.

Callender, C. (2008). The common now. *Philosophical Issues, 18*, 339–361.

Chalmers, D. J. (2002). Does conceivability entail possibility? In T. S. Gendler & J. Hawthorne (Eds.), *Conceivability and possibility* (pp. 145–200). Oxford: Oxford University Press.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Coffa, A. (1993). *To the Vienna station*. Cambridge: Cambridge University Press.

Craig, W. L. (2001). *Time and the metaphysics of relativity*. Dordrecht: Kluwer.

Deutsch, D. (1999). Quantum theory of probability and decisions. *Proceedings of the Royal Society of London, A455*, 3129–3137.

Dieks, D. (2006a). Becoming, relativity and locality. In D. Dieks (Ed.), *The ontology of spacetime* (pp. 157–176). Amsterdam: Elsevier.

Dieks, D. (Ed.). (2006b). *The ontology of spacetime*. Amsterdam: Elsevier.

DiSalle, R. (2006). *Understanding spacetime*. Cambridge: Cambridge University Press.

Dolev, Y. (2006). How to square a non-localized present with special relativity. In D. Dieks (Ed.), *The ontology of spacetime* (pp. 177–190). Amsterdam: Elsevier.

Dorato, M. (2006a). The irrelevance of the presentism/eternalism debate for the ontology of Minkowski spacetime. In D. Dieks (Ed.), *The ontology of spacetime* (pp. 93–109). Amsterdam: Elsevier.

Dorato, M. (2006b). Absolute becoming, relational becoming and the arrow of time: Some non conventional remarks on the relationship between physics and metaphysics. *Studies in History and Philosophy of Modern Physics, 37*, 559–576.

Dorato, M. (2012). Presentism/eternalism and endurantism/perdurantism: Why the unsubstantiality of the first debate implies that of the second. *Philosophia Naturalis, 49*(1), 25–41.

Eddington, A. S. (1928). *The nature of the physical world*. London: McMillan.

Fine, A. (1989). Do correlations need to be explained? In J. Cushing & E. McMullin (Eds.), *Philosophical consequences of quantum theory* (pp. 175–194). Notre Dame: Notre Dame University Press.

French, S., & McKenzie, K. (2012). Thinking outside the (tool) box: Towards a more productive engagement between metaphysics and philosophy of physics. In A. Cei & M. Dorato (Eds.), *Physics meet philosophy and bridging two gaps*. In honor of Michael Dummett. *European Journal of Analytic Philosophy, 8*(1), 42–59.

Giere, R. (1988). *Explaining science*. Chicago: University of Chicago Press.

Ladyman, J., & Ross, D. (2007). *Everything must go: Metaphysics naturalized*. Oxford: Oxford University Press.

Lange, M. (2002). *An introduction to the philosophy of physics*. Oxford: Blackwell.

Lowe, E. J. (1998). *The possibility of metaphysics*. Oxford: Clarendon Press.

Lowe, E. J. (2011). The rationality of metaphysics. *Synthese, 178*, 99–109.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve belief about the motion of object. *Science, 210*(4474), 1139–1141.

Morganti, M. (2013). Scienza e metafisica. Per un realismo costruttivo. *Sistemi intelligenti Anno, 25*(1), 67–82.

Psillos, S. (2011). Living with the abstract: Realism and models. *Synthese, 180*(1), 3–17.

Savitt, S. (2002). On absolute becoming and the myth of passage. In C. Callender (Ed.), *Time, reality & experience* (pp. 153–167). Cambridge: Cambridge University Press.

Savitt, S. (2006). Presentism and eternalism in perspective. In D. Dieks (Ed.), *The ontology of spacetime* (pp. 111–127). Amsterdam: Elsevier.

Sellars, W. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.), *Frontiers of science and philosophy* (pp. 35–78). Pittsburgh: University of Pittsburgh Press. Reprinted in *Science, perception and reality* (1963). London: Routledge.

Van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon Press.

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.

How (Not) To Be a Humean Structuralist

Kerry McKenzie

Abstract While the idea that the structures of ontic structural realism should be understood as in some sense 'modal' has been referred to many times, comparatively little has been said regarding how exactly that modality should be understood. However, Lyre has recently defended the idea that a Humean interpretation of structures is possible by understanding them to be composed of 'categorical' properties and relations. In this paper I raise some objections to deferring to the notion of categorical properties to articulate a modal interpretation of structures, and gesture towards an alternative means of expressing a Humean form of structuralism.

1 Introduction

The idea that the structures of ontic structural realism are to be understood as in some sense 'modal' has often been gestured at, but how exactly that modality is to be understood has received comparably little by way of discussion. Recently, however, Michael Esfeld and Holger Lyre have both articulated explicitly modal interpretations of structures – though they have very different stances on what they take the modal commitments of structuralism to be. Esfeld for example 'appl[ies] the debate about causal vs. categorical properties in analytic metaphysics to ontic structural realism' in order to develop a non-Humean account of structures, where their non-Humean nature is secured by the fact that the relations comprising them are understood to be irreducibly causal, or *essentially dispositional* (Esfeld 2009, p. 179). Lyre by contrast adopts a view in which the properties and relations that

K. McKenzie (✉)
Department of Philosophy, University of Leeds, LS2 9JT, Leeds, UK
e-mail: mckenzie_kerry@ymail.com

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_25,

comprise the relevant structures are understood to be *categorical* in nature, and takes it that a Humean perspective on structures results from understanding them in such terms (Lyre 2010).[1]

While the modal interpretations of structures offered by Esfeld and Lyre are diametrically opposed, the *strategy adopted* to articulate these interpretations is the same in both cases. What is assumed by each author is a modal distinction applying to properties that is familiar from analytic metaphysics – namely that between *essentially dispositional* and *categorical* properties – which is then appealed to in order to ground distinct modal interpretations of the relevant structures.[2] This strategy exactly parallels that which is adopted in (what I will call) the 'canonical' debate over laws of nature, in which Humean and non-Humean interpretations of laws are grounded in opposed modal accounts of the nature of fundamental properties. That such a parallel exists is, of course, in many ways unsurprising, given that laws themselves are often taken to be paradigmatic examples of structures in physics.[3]

While this strategy for articulating modality may seem inevitable and natural, I want to argue here that it is nonetheless problematic for structuralists to adopt it. Ontic structuralism is after all a resolutely *naturalistic* thesis, and one that ultimately aims to give an account of the fundamental nature of reality; as such, and as I will argue, it is entirely unclear that structuralists can blithely appeal to a modal conception of properties that has been incubated in the context of analytic metaphysics, given that the latter is often charged with being wedded to *too classical* a picture of reality to be of service in fundamental regimes. My objections will be directed in this instance toward the uncritical invocation of, in particular, *categorical* properties in the context of fundamental physics, and thus toward Lyre's account of modality in structuralism which is predicated upon it. I stress, however, that in so doing I am *not* thereby defending the rival non-Humean account, such as that offered by Esfeld: since I am suspicious not just of the notion of categorical properties, but of the essentially dispositional/categorical distinction itself, for me it is a case of 'a curse on both your houses' insofar as the debate over modality is constructed upon it.

In what follows, I will focus on the fundamental *kind* properties, and my argument will proceed in two stages. I will argue that

(i) The modal metaphysics standardly associated with categorical properties assumes an account of natural law that not appropriate for elucidating fundamental properties; and

(ii) If we move to a more realistic account of fundamental laws, and if we take the QM formalism seriously, it isn't clear that there is any place for categorical properties in our metaphysics – at least not as standardly conceived.[4]

[1] As Lyre writes, 'A proper Humean perspective on structural realism is to demand categorical structures and to dismiss mysterious modalities' (ibid., p. 10).

[2] I will subsume relations under the term 'properties'.

[3] One need think only of the structuralist discussions of Fresnel's and Maxwell's equations.

[4] Whether there is a different but sufficiently analogical way of understanding categorical properties that does not fall victim to the objections I raise is an interesting question, but not something I can discuss here.

If we want to articulate a Humean version of structuralism, then, I think we should try to find another strategy that does not appeal to the concept of categorical properties, and I will hint at the shape that such a strategy might take at the end. For now, however, I will outline how I understand the canonical account of laws, properties and modality in which the notion of categorical properties was developed. Once that is in place, I will be able to articulate some of the problems that I perceive in the act of appealing, in the fundamental physics context, to categorical properties so conceived.

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. On the one hand, we have *non-Humean* accounts in which laws are taken to consist of metaphysically necessary connections between properties. In the contemporary literature, such accounts are associated with authors such as Bird and Ellis (see e.g. Bird 2007; Ellis 2001). On the other hand, we have *Humean* accounts in which laws consist of metaphysically contingent connections between properties. Such accounts are primarily associated at present with authors such as Armstrong and Loewer (see e.g. Armstrong 1997; Loewer 1996).[5]

Each of these modal accounts of laws – just as with Esfeld's and Lyre's accounts of structures – is typically *grounded* in a prior modal conception of properties. Non-Humeans 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, on this view a given species of fundamental particle, defined by a given set of fundamental properties, can act in accordance with *one and only one* law across different possible worlds. It is thus this modal conception of properties that non-Humeans typically take to account for the fact that the laws are metaphysically necessary. By contrast, Humeans reject this view of fundamental physical properties, and as such also the idea that the kinds that instantiate such properties bring in their wake a unique law. They rather endorse an opposing view of properties in which they are deemed 'categorical' in nature, and it is this categorical conception of properties that is taken to underwrite the idea that a given kind of particle could behave differently.

However, and while what exactly is involved in the concept of an essentially dispositional property has been discussed at length in many places, 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). And of course, without *some* such specification the precise connection between categorical properties and the contingentist interpretation of

[5] Since I am drawing the distinction between the two positions in terms of Hume's dictum and not primitive modality, I (like Bird) place Armstrong's analysis in the Humean category.

laws can only remain murky. One can, however, find a variety of strategies that are used to at least gesture at what is intended by this designation. One finds categorical properties characterized, for example,

(i) 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 op. cit., pp. 69, 80); or alternatively
(ii) In explicitly *nomological* terms, as those properties that are 'free of nomic commitments' (Carroll 1994, p. 8), or as those that do not 'necessarily involve laws' (Loewer op. cit., p. 200); or sometimes
(iii) 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).

There thus seem to be a number of ways of approaching what is meant by a categorical property. Greater variety does not equate with greater clarity, however, and it would be nice if what is meant by 'categorical' in this context could be sharpened up. A strategy frequently adopted to convey more precisely what it is that is meant is that of simply *conveying by example* the implications of such properties for the laws of nature. So for instance, it is often cited that on this 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).[6]

While the specific examples offered of alternative laws are typically rather conservative in how they differ from actual laws – consisting in these cases just of a sign flip and a unit increase of power respectively – such discussions nonetheless tend to be silent on what *principles govern* how the actual laws may be tinkered

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

with so as to generate acceptable other-worldly alternatives. But without some such statement, the exact relationship between categorical properties and possible variation in laws – and hence the concept of categorical properties itself is still problematically hazy. Perhaps we should take it – since such properties are regarded as 'free of nomic commitments' – that it simply goes without saying that there *are* no such principles (or at least no non-trivial ones). But if *that* is the case, then we can improve upon this strategy of conveying by example what is meant by 'categorical' by moving to a more general – and thus more definitive – characterization in the following way.

Recall that the example that we just looked at was that of Coulomb's law. This law is a paradigmatic example of a *classical* law, and 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 conception of laws that is in play in the contemporary debate on laws of nature is *not* the old $\forall x(Fx \rightarrow Gx)$-type formulation that was 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... Only if we can give some plausible account of functional laws... do we have a theory of lawhood that can be taken really seriously (Armstrong 1997, p. 242).

Assuming such an account of fundamental laws, then, we can better formalize what is at issue between the two camps in the canonical debate over their modal status. 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 A to D. Non-Humeans will then hold that, since the fundamental properties are *essentially dispositional*, then

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

and in particular that

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

where $f' \neq f$. 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 properties. By contrast, Humeans will hold that

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

and in particular that

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

As mooted above, if properties are categorical then it seems there should be no non-trivial constraints on the form of the laws that any such properties feature in, and hence no non-trivial constraints on the choice of f'.[7] 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 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'.

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 that sorts them into 'categorical' and 'essentially dispositional' properties – where I take the former to be most perspicuously defined as above – and that modal distinction between properties is used to ground a corresponding modal distinction between laws. The laws of principal interest are the *fundamental* laws, where these are assumed to have a functional structure. But when the terms of the debate are stated in that way, it becomes immediately evident that there is a very basic problem afoot in it. That problem is that this debate over laws in analytic metaphysics purports to describe fundamental laws and properties, and thus capture the metaphysics of fundamental physics; *but fundamental physics properties do not obey functional laws*![8] The reason for this, of course, is that fundamental properties and their laws must be understood within the framework of quantum theory, and quantum-theoretic laws are not – and cannot be – of functional form. But since categorical properties have been *defined* in terms of the relationship they bear to functional laws, we need to consider whether any fundamental property can

[7] 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 occurence 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.

[8] Or, if we count charge as a fundamental property (which is controversial), at least not in its most 'fundamental guise'.

properly be regarded as such when the latter are out of the picture.[9] Let me therefore now consider whether any fundamental properties may be regarded as categorical in the context of quantum theory, and thus whether appeals to the notion of categorical properties may still be made in that context to ground a Humean interpretation of laws and other structures. As above, I will continue to focus on the fundamental kind properties.[10]

3 Laws and Properties after Quantum Mechanics

While ideally I would directly discuss laws in quantum field theory, I will focus just on the representation of laws in quantum particle mechanics and recount only their basic features.[11] The nearest thing that we have in quantum particle mechanics to the functional template for laws in classical physics is of course the Schrodinger equation:

$$i\hbar\frac{\partial|\psi\rangle}{\partial t} = H|\psi\rangle.$$

Expressed a little more fully, laws of the Schrodinger form are statements

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

[9] Since essentially dispositional properties are characterized in terms of their entailment of such laws, analogous problems will apply to them.

[10] Since state-dependent properties 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.

[11] A referee has rightly pointed out that it may not be appropriate to argue for metaphysical conclusions by focusing only on the textbook formalism, as I do here, without taking into account the different interpretations of that formalism. In particular, they argue that if one does not accept that the Hamiltonian corresponds to a fundamental local beable, or that commutation relations are a guide to the fundamental ontology of physics, then one can understand the Hamiltonian as a mere compendium of correlations between events involving categorical properties understood in the third, spatiotemporal sense in the list above. However, even *if* that is the case, I do not think that it undermines the present concerns. For one thing, it remains that owing to the different formal concepts of laws and their relationships to properties, the first two conceptions of categorical properties listed above – conceptions which are (implicitly or explicitly) expressed in nomological terms – cannot but come under pressure by the considerations adduced here; since the topic of this paper is simply that one cannot uncritically export the concept of categorical properties that was developed in a classical ('functional law') framework into the current context, the mere fact that *some* renderings of such properties are ruled out is enough to make the point. Secondly, however, Lyre is trying to develop a specifically structuralist form of Humeanism, in which commutation relations – especially those involved in the definition of symmetry structures – emphatically *are* regarded as the fundamental ontology (see Lyre op. cit., pp. 2, 11), though not as 'local' but as 'global' beables (ibid., p. 11). Thus these formal considerations most certainly *are* sufficient to generate problems, for this variant of Humeanism at the very least.

where H_α denotes a specific Hamiltonian and the n_i denote the properties that identify the kind, or kinds, of particle involved.[12] These Hamiltonians describe both how a single particle's states evolve through its Hilbert space, and also contain all the information about a particle's *interactions* with other systems. For example, the quantity

$$\langle (n, \pi^+)|H_S|(p, \pi^-)\rangle$$

yields the probability that two different particle kinds, a negative pion and a proton, will interact through the strong interaction to produce a positive pion and a neutron.

These facts are of course utterly elementary, but they have immediate and non-trivial implications on whether the fundamental kind properties may be properly deemed categorical. Suppose, for example, that we have particle kind defined by a set of determinate properties $\{n_i\}$ acting in accordance with a law of the above form. Talk of a given kind of particle evolving in time presupposes that the set of properties defining that kind are preserved through time, and hence are conserved by the corresponding Hamiltonian. Within the formalism of quantum mechanics, then, the kind structure of a given world is defined in terms of those properties whose operators *commute* with at least one Hamiltonian operating in that world. Thus to claim that any such world contains a kind property n_i requires us commit to there being a law in that world involving a Hamiltonian H_α such that $[H_\alpha, N_i] = 0$, where N_i is the operator corresponding to n_i.[13]

Talk of kind properties in quantum mechanics thus brings in its train the demand that (at least some of) the Hamiltonians operating in a world in which the relevant kinds exist have certain structural features – that is, that they satisfy commutation relations with the operators corresponding to those kinds. This demand, however, represents a *non-trivial constraint* on the structure of those Hamiltonians, and hence on the form of the laws governing those kinds – non-trivial in the sense that it can fail.[14] But we already saw that in the canonical account, this was something that categorical properties *did not do*. In that account, a categorical property was one that was 'independent of its nomic role', and that seemed to imply that there were no (non-trivial) constraints on the *mathematical form* – in that case, the functional form – of the laws that particles with that property could partake in.

We can thus already see that there is a difficulty with blithely importing the concept of categorical properties – a concept that was incubated in analytic metaphysics against a background of classical physics – into the metaphysics of

[12] Some state-dependent variables x_i should also be included in the characterization of the state, but my focus here is just on kind properties.

[13] Of course, analogous considerations apply in classical Hamiltonian mechanics as well; so much the worse, in my opinion, for the discussion of modality in analytic metaphysics. Nonetheless, some of the considerations I adduce below are intrinsically quantum mechanical.

[14] For example, the parity operator – against all expectation – was found not to commute with the weak-interaction Hamiltonian. It is nonetheless still used as a classificatory device since it is conserved by other interactions.

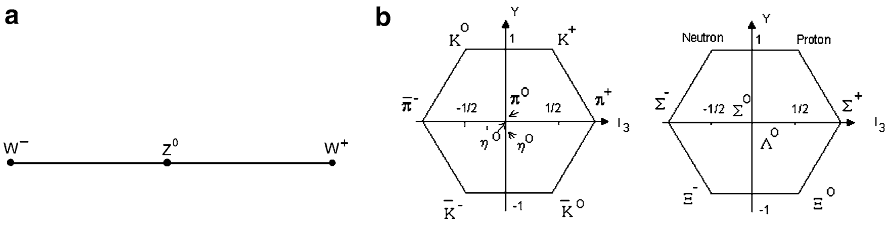


Fig. 1 Actual particle multiplets (**a**) SU(2) triplet of weak bosons. (**b**) SU(3) octets of hadrons

fundamental physics. This difficulty is on account of the constraints imposed by the commutation requirements on the laws in any world containing a given kind structure. Now, to see just how non-trivial commutation constraints can be, one need only consider the impact that *symmetries* can have on any realistic discussion of nomological modality. To say that a law in quantum mechanics possesses a symmetry is to say that there is a set of operators U_i such that (i) the U_i form a group (in the mathematical sense) and (ii) for all U_i, $[H, U_i] = 0$, where H is the Hamiltonian corresponding to that law. The presence of a symmetry has important consequences for the solutions of the Schrodinger 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 the n_i again represent a set of determinate properties defining some kind, and the n_j a different set of determinate properties *but of the same determinables* as those that define the first. Thus where there are symmetries of the laws, there are *families of particles* that obey those laws with the same energy (hence mass), but different *determinate* values of the same *determinable* properties. Such of families of particles are called 'multiplets'.

As it turns out, 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 interaction, and the SU(3) symmetry of the strong interaction. That of course means that the particles that populate this world themselves fall into such multiplets. We have for example in Fig. 1a the triplet of the weak bosons, corresponding to the 3 representation of the SU(2) symmetry, and in Fig. 1b some of the hadrons comprising SU(3) multiplets (the gluons do so likewise).

These diagrams represent elegant facts about the fundamental structure of the actual world, but their principal relevance for topic at hand may be seen once one recalls that debates over the modal status of laws are often framed in terms of *duplicates*. We know that non-Humeans hold that otherwordly duplicates of actual particles cannot act in accordance with different laws; as Bird puts it, 'If the particles and fields are the same in the two worlds then they instantiate the same [essentially dispositional properties] and thus give rise to identical laws' (Bird 2007, 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) – and as I have argued, seem to be committed, through their commitment to categorical properties, to their being subject to *arbitarily* different laws. What, then, is the situation here? Can otherworldly duplicates of the actual particles, which as we know occur in *multiplets*, obey arbitrarily different laws?

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. But it is easy to show that each such weight diagram corresponds to *one and only one algebra*. What that informs us of in turn is that, wherever in possibility space duplicates of these actual particles are instantiated, the laws that hold there *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. Indeed, one often hears particle physicists recite the adage that 'symmetries dictate laws': to the extent that is correct, then it follows that duplicates of actual particles must obey a *unique* law wherever it is that they are instantiated in possibility space.[15] Such a view of laws as metaphysically necessary is of course associated in the canonical picture with the non-Humean view – a view that in turn is based on the rejection of categorical properties. How, then, can one possibly maintain in particle physics a view of the fundamental properties as 'free of nomic commitments', and a corresponding Humean stance toward laws?

Before expanding on that question, I want to take a brief segway to raise a point that gestures, in my mind, to just how radically the debate over nomological modality may have to change if it is to be appropriately reflective of realistic fundamental physics. As just pointed out, considerations of the bearing of the mathematics of symmetry on the question of how duplicates of actual particles can behave led us to something close to the metaphysical necessity traditionally associated with the non-Humean camp; *how* close will be a function of how seriously we take the (problematic) adage that 'symmetries dictate laws'. In the canonical account, that uniqueness was grounded in a prior assumption about modal nature of properties – namely, that they are 'essentially dispositional'; *here*, however, the restrictions on the laws that any given set of particles may accord with *was derived just through the mathematics of symmetry*, applied in the QM framework. But then what exactly the *conflict* with Humeanism consists in is not clear, since Humeans – while suspicious of general metaphysical necessities – are of course perfectly happy to sanction *mathematico-logical* necessities, and hence presumably necessities such as these. What we thus seem to be contemplating is at least the *coherence* of a view in which a broadly Humean metaphysics may be combined with a view of laws as metaphysically necessary, since the latter issued just from the relevant mathematics applied against the backdrop of a quantum representation of laws. Since there is simply no analogue of this in the canonical debate over nomological modality, that in turn suggests that its basic terms may have to be radically revised if it is to be relevant to contemporary fundamental physics – revisions that may extend so far as to undermine the basic Humean–non-Humean

[15] Such a claim is often made in the context of gauge symmetries.

dichotomy that defines the basic structure of such debates. And that in turn, of course, should make us yet more suspicious about Lyre's (and Esfeld's) strategy of borrowing concepts developed within that debate to articulate a modal metaphysics for structuralism.

4 Conclusion

To finish, then, let me review the impact of the above considerations on the view, deferred to by Lyre, that the fundamental properties are categorical. We have seen that the attribution of kind properties post quantum mechanics brings in its wake non-trivial – and sometimes *highly* non-trivial – constraints on the form of laws. May we thus regard the fundamental kind properties as being categorical in nature? In other words, and going back to list above, may we consider the fundamental kind properties

- As properties that don't 'look outward' to interactions? The answer is that *it doesn't seem so,* since constraints on Hamiltonians are *ipso facto* constraints on interactions.
- As properties that are 'free of nomic commitments'? Again, *it doesn't seem so,* since one is commited within this framework to the satisfaction of relevant (and non-trivial) commutation relations wherever one defines a kind structure.
- As properties 'whose instantiation has no metaphysical implications concerning the instantiation of fundamental properties elsewhere and elsewhen'? Again, *it doesn't seem so*. It is after all standard practice to represent laws as *global* entities, as properties of worlds themselves: the constraints on the laws implied by the instantiation of a kind of particle in that world – however non-localized, short of being *globally* instantiated, that particle may be – are therefore implications for parts of spacetime that it does not inhabit.

What I hope this all this has shown is that if we want to articulate a Humean metaphysics of fundamental physics – at least if we take the formalism seriously – then we cannot simply defer to the notion of categorical properties in order to do so. As I have hinted, however, I think that the problems with the typical metaphysical discussions of laws, properties and modality are more general, and go deeper, than any specific problem with the notion of categorical properties per se. But since I suspect that it may be a lack of attention to the *mathematics* of physics that lies at the root of many of these problems, and since a large part of structuralism has consisted of trying to better integrate the mathematics of physics with its metaphysics, perhaps a close study of the modal commitments of structuralism is *exactly the right place to start* if we want to move beyond the traditional debates. Thus while I believe that the work of Lyre, and Esfeld, is flawed as it stands, it may turn out to be a highly valuable springboard for a more physically engaged study of naturalistic modal metaphysics.

References

Armstrong, D. M. (1997). *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.

Carroll, J. (1994). *Laws of nature*. Cambridge: Cambridge University Press.

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

Esfeld, M. (2009). The modal nature of structures in ontic structural realism. *International Studies in the Philosophy of Science, 23*, 179–194.

Lewis, D. (1986). *On the plurality of worlds*. Oxford: Wiley-Blackwell.

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

Lyre, H. (2010). Humean perspectives on structural realism. In F. Stadler (Ed.), *The present situation in the philosophy of science* (pp. 381–397). Dordrecht: Springer. (References taken from PhilSci Archive preprint 4574)

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

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

Part VIII
Philosophy of the Physical Sciences: Philosophy of Chemistry

What Does Hydrogen Bonding Say About the Nature of the Chemical Bond?

Paul Needham

Abstract The status of the chemical bond has long been a controversial issue with the increasing distance between quantum chemists' theoretical understanding of molecular stability and the ideas of experimental chemists. Some aspects of the development of the concept of a hydrogen bond are discussed with a view to assessing its import on the general question.

1 Introduction

Coulson's famous charge that "a chemical bond is not a real thing: it does not exist: no-one has ever seen it, no-one ever can. It is a figment of our own imagination" (1955, p. 2084) bodes ill for any hopes of pinning down a "nature" of the chemical bond. The present paper considers how specifically hydrogen bonding (H-bonding) adds further grist to the mill. Whether H-bonds are chemical bonds is not entirely clear-cut, however, so the bearing of this specific topic on the general theme might be disputed. But I argue that H-bonds are as good chemical bonds as any. This does illustrate a general problem of drawing clear boundaries when it comes to bonds, but the lesson to be drawn isn't encouraging for anyone looking for a nature.[1]

H-bonding has been studied ever more intensively in the course of the last century as its importance in many branches of chemistry has become more apparent. Throughout this time, the question of how the concept should be delimited has been

[1] Limitations of space preclude a historical overview of the development of the idea of a chemical bond. A good discussion with many useful references is Sutcliffe (1996).

P. Needham (✉)
Department of Philosophy, University of Stockholm, SE-106 91 Stockholm, Sweden
e-mail: paul.needham@philosophy.su.se

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_26,

a lively issue. Recently, the International Union of Pure and Applied Chemistry (IUPAC) set up a Task Group to reconsider the definition of the H-bond. The group's report was published in 2010 after an unusually thorough peer review procedure involving 25 reviewers. Their proposal for the revised definition is difficult to reproduce here short of reproducing the entire paper. There is a shortish formulation:

> The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule in which there is evidence of bond formation. (Arunan et al. 2010, p. 12)

But this is supplemented with two lists clarifying the meaning of "there is evidence of bond formation", one giving six criteria indicating what counts as evidence and the other describing six characteristics of H-bonds, and all items are qualified by substantial footnotes. The authors say these lists are neither strictly universal nor final, but open to further qualification, implying that we are not dealing with a definition in the strict sense of the term. Considered in the light of the historical development of the concept, the presupposition of the Task Group's endeavour, that a unified concept seems to be emerging, seems reasonable. But H-bonding cannot be attributed to a "single physical force" (Arunan et al. 2010, p. 5).

The historical development of the concept can only be hinted at within the confines of this short article, which falls short of a review of even the major points that would be required to justify the Task Group's presupposition. I will outline why the first substantial evidence of H-bonding in the first decades of the twentieth century provided little indication that this kind of interaction should be counted as a chemical bond. Subsequent investigations reversed this picture, although not without cutting across previously established distinctions. More recent theoretical calculations provide a general picture explaining the broad features of H-bonding and saying something about the distinctive character of H-bonding, although exploration of the details does not point in the direction of a single underlying physical force.

2 Hydrogen "Bonds" are Bonds

The establishment of the law of constant proportions at the beginning of the nineteenth century provided a criterion for the occurrence of chemical combination, distinguishing compounds, which usually presented themselves as homogeneous matter, from solutions – also homogeneous mixtures understood to be purely mechanical mixtures. As thermodynamics was shown to be sensitive to distinctions of substance later in the century, providing the first general theory of chemical combination based on the thermodynamic potentials, it facilitated the detailed study of solutions. Chemists were able to characterise the idea of a "purely mechanical" mixture as an ideal solution (whose greater stability compared with the separated components is entirely due to the entropy of mixing). Ideal solution behaviour was only approximated under special conditions for a restricted class of mixtures, however. Non-ideal behaviour was more commonly observed in mixtures homogeneous over a continuous range of proportions of their components. It became clear that solutions typically exhibit some kind of interaction between their components, and display the same general kind of reduction in thermodynamic potential compared to the isolated ingredients as do compounds. Chemical combination in the broad sense of the term was not restricted to compounds.

Early in the twentieth century, ideas about microstructure began to do some work in chemistry and the chemical combination holding elements together in compounds was now called chemical bonding. Following Gilbert Lewis, a broad distinction was drawn between ionic and covalent compounds according as their elemental constituents were held together by ionic or covalent bonds. Some substances in the category of covalent compounds are molecular, i.e. macroscopic quantities of them comprise collections of particles of a single kind called molecules. Molecules are held to one another, fairly rigidly in the solid state and less tenaciously in the liquid state, by *intermolecular* forces, distinguished from the much stronger *intramolecular* forces holding the atoms together by chemical bonds and giving rise to the structure of the molecule. It is the intermolecular forces that were held responsible for the non-ideal behaviour of solutions of molecular substances.

The concept of H-bonding, if not the name (which was coined by Pauling in the 1930s), emerged around 1920 from Lewis's laboratory to explain thermodynamic data such as the abnormally high melting points, boiling points and latent heats of vaporisation of the hydrides of the first row elements nitrogen, oxygen and fluorine compared with trends followed by hydrides of other elements in the groups of the periodic table to which these elements belong. Although stronger than typical intermolecular van der Waals and London forces, the intermolecular force mediated by a hydrogen bridge that was postulated to explain these abnormalities was considerably weaker than typical intramolecular covalent bonds. It was strongest in hydrogen fluoride, although the abnormal effect was most marked in the case of water. This was explained by greater density of H-bridge bonding in water because of the ability of water to sustain a three-dimensional array of H-bridges in which each water molecule was the locus of four such H-bridges (as allowed by the postulate of Latimer and Rodebush described below).

At this stage, then, the new hydrogen-mediated interaction seemed less like a covalent bond and more like the other intermolecular forces that were contrasted with bonds because they were considerably weaker and not intramolecular. But this sharp contrast was muddied as two new features in particular came to light.

The first of these was first advanced by Linus Pauling, who after having developed an influential theoretical account of H-bonding in the 1930s, came to see in the ensuing decades how the concept could be used to explain aspects of the structure of large biological macromolecules such as carbohydrates and proteins which outreached the devices traditionally used in the structural conception of organic molecules to explain isomerism. The so-called secondary structure of proteins, which deals with how the covalently linked chains of peptides (formed by condensation of two amino acids) are folded, is held in place by hydrogen bonding.

This accounts for different stereoconformations of macromolecules with the same primary structure (succession of amino acids of the polypeptide “backbone”) having very different chemical properties. Denaturation – the process of destroying the secondary structure of a protein induced by changing the pH or heating – transforms the protein into a biologically inert polypeptide. Watson and Crick were subsequently able to use the idea to explain how the double helix of DNA is held in place.

Here we see H-bonding acting as an *intramolecular* force, contributing to the internal structure of molecules, rather than merely holding different molecules together, and determining chemical properties. This cuts across the previously accepted distinction between bonds and other forms of interaction.

The second feature distinguishing H-bonds from intermolecular forces of the van der Waals and London types is their directionality. This feature is already apparent in the intramolecular H-bonding, which fixes the relative orientation of submolecular units in the secondary structure of macromolecules. But it is a general characteristic of H-bonding, and one of the features the IUPAC Task Group emphasised in their characterisation. A nice illustration is provided by the comparison of hydrogen sulphide and water. In the early days, H-bonding was thought to be restricted to hydrogen bound to N, O and F. But with the realisation that it is considerably more widespread, the condition of being more electronegative than hydrogen is seen as raising the prospect of being a source of H-bonding. Sulphur is more electronegative than hydrogen, with a value of 2.589 compared with hydrogen’s 2.3 on the revised Pauling scale and so the question arises whether H_2S is a possible source of H-bonding. X-ray crystallography shows, however, that when H_2S freezes at –60 °C, molecules in the crystal structure have 12 neighbours, which is the most efficient packing of the nearly spherical H_2S molecules typical of the non-directional bonding in ionic crystals. In ice, on the other hand, each H_2O molecule has just four neighbours, oriented tetrahedrally around the central H_2O molecule. Consequently, ice is not as dense as it would be if the H_2O molecules had a close-packed structure like that of H_2S.

H-bonds are beginning to look more like covalent bonds. They occur as intramolecular forces contributing to the shape of molecules, and they have the general feature of orienting the units they bind along specific directions in space. Geometry has become especially important since neutron diffraction in the last few decades has made possible reliable localisation of H-bonded protons in H-bonds of the kind X–H • • • Y, so complementing X–Y distances determined by old techniques with X–H and H–Y distances, enabling the determination of the XHY angle. What sets H-bonds apart from typical covalent bonds would seem to be their relatively low bond energies. But alongside the growing recognition that H-bonding is far more widespread than was previously thought came the discovery that H-bonds are not all as weak as was thought in the early days. H-bonds are now roughly categorised as weak, moderate and strong, and with this range of bond energies, the sharp distinction between H-bonds and covalent bonds in respect of bond energy has been eroded. Bond energies are one of the major features theorists hope to calculate by way of confirming theories of the bonds in question, but it wasn’t the theoreticians who gave the lead in expanding the range of recognised H-bond energies.

3 Shortcomings of Theory

Covalent bonds are, in accordance with Lewis's model, typically electron pairs linking two atoms in a molecule. H-bonds are not like this, but rather three-centre, four-electron bonds, written X–H•••Y, where X–H unit is called the H-bond *donor* and Y is the H- *acceptor*. This was the original conception of Latimer and Rodebush, who proposed the concept while working in Lewis's laboratory to explain the highly associated character of liquids like water and HF, the small basicity of ammonium hydroxide and dimerisation of acetic acid. According to their understanding of what they only hesitatingly called a bond,

> a free pair of electrons on one water molecule might be able to exert sufficient force on a hydrogen held by a pair of electrons on another water molecule to bind the two molecules together. ... Such an explanation amounts to saying that the hydrogen nucleus held between two octets constitutes a weak 'bond'. (Latimer and Rodebush 1920, p. 1431)

Chemists veered from this general conception some time later under the pressure of accommodating a broader range of phenomena, seeking a less theoretical, more practically and experimentally oriented characterisation. But they returned to it in more recent times.

Pauling's account of H-bonding, summarised in the first edition of *The Nature of the Chemical Bond* (1939), substantially influenced all work in the field for 2 decades. He understood H-bonding largely as a weak bond calling for an electrostatic explanation since

> the hydrogen atom, with only one stable orbital (the 1s orbital), can form only one covalent bond, ... [so] the *hydrogen bond* is largely ionic in character, and ... formed only between the most electronegative atoms. (Pauling 1960, p. 449)

H-bond energies lie "in most cases in the range 2–10 kcal/mol" (p. 449). Pauling did recognise the occurrence of strong, symmetric H-bonds, e.g. $[F\bullet\bullet\bullet H\bullet\bullet\bullet F]^{-}$ in KHF_2, noting that the distance between the fluorine atoms, 2.26 Å, is unexpectedly short (less than the sum of the van der Waals radii). Here the covalent character of the bond predominates, which he explained by invoking resonance between two valence bond covalent forms X–H•••X and X•••H–X, claiming that "the hydrogen atom in the $[HF_2]^{-}$ ion lies midway between the two fluorine atoms and may be considered to form a half-bond with each" (1960, p. 484). Nevertheless, such cases where the covalent character of the bond predominates were extremely rare, and he regarded them as exceptions which "can be reasonably neglected in the treatment of the much more copious H-bonds of normal strength" (1960, p. 485). Presumably Pauling was seeking a simple nature; otherwise, this attitude is difficult to understand.

This persuaded many workers in field that covalency was not an important factor to be reckoned with. As it became apparent that H-bonding is much more widespread than Pauling realised, however, and not confined to most electronegative elements N, O and F, chemists veered from the original conception. A less theoretical approach was heralded by the appearance of Pimentel and McClellan's widely respected book *The Hydrogen Bond* in 1960. According to these authors, a hydrogen bond A–H•••B exists where

> (a) there is evidence of bond formation (association or chelation), (b) there is evidence that this new bond linking A–H and B specifically involves a hydrogen atom already bonded to A. (Pimentel and McClellan 1960, p. 6)

Boiling point and freezing point modification would be examples of (a), and shift of infra red and Ramen A–H stretching frequencies, ^{1}H NMR chemical shift of the proton or the A•••B and AH•••B distances determined by X-ray and neutron diffraction becoming much shorter than the sum of their van der Waals radii would be examples of (b), distinguishing H-bonds from other types of associative interaction.

Pimentel and McClellan's characterisation has been praised as a pragmatic definition, paving the way for the introduction of the tripartite classification of H-bonds as strong, moderate and weak. But it came to be seen as too broad. A case in point is the B–H–B bond occurring in boranes, which is an H-bond by Pimentel and McClellan's lights (as they explicitly argue). Although a three-centre bond, it involves only two electrons, whereas on Latimer and Rodebush's conception two pairs are required, making the H-bond a three-centre, *four* electron bond. Further, boron is less electronegative than hydrogen (2.051 compared to 2.3), which raises the issue of the polarity of H-bonds. Pimentel and McClellan's definition makes no ruling on this point, but Pauling's conception involving a hydrogen between electronegative atoms would mean that the polarity must be $X^{\delta-}$–$H^{\delta+}$•••$Y^{\delta-}$, and not $X^{\delta+}$–$H^{\delta-}$•••$Y^{\delta+}$. Arunan et al. (2010, p. 7) make a similar point about halogen-bonded complexes which are H-bonded by Pimentel and McClellan's definition. The HF•••ClF complex, for example, was first thought to involve a hydrogen bond, ClF•••HF, but is now recognised to involve a chlorine bond more appropriately represented as HF•••ClF. H may well be bonded to two atoms in complexes such as LiH•••$ClCF_3$, where the hydride acts as a halogen-bond acceptor. This satisfies Pimentel and McClellan's definition, but Arunan et al. argue it should be classified as chlorine-bonded, and the notion of an H-bond restricted to cases where donor atom is more electronegative than H.

4 Aspects of the MO Account

Doubts about Pauling's understanding of H-bonding as a purely electrostatic interaction began to set in by the 1950s. In the 2nd ed. of *Valence* (1961), Coulson was urging that "[t]he fact that the energy of a hydrogen bond is essentially electrostatic does not mean that no resonance effects occur" and "no completely satisfactory account can be given without including several factors not normally required in discussing conventional chemical bonds" (1961, p. 353). The factors he had in mind included delocalisation, i.e. charge transfer effects associated with partial covalent character, repulsive overlap of charge clouds on nonbonded atoms and dispersion forces, which he incorporated within a Pauling-style valence bond (VB) treatment. But the molecular orbital (MO) method came into favour in the course of the 1960s, which depicted MOs as delocalised over whole molecules, unlike the VB treatment of bonds which corresponded with the more localised conceptions of bonds featuring in general chemistry textbooks. A crude account in MO terms depicts an H-bond X–H•••Y as a modification of the isolated X–H donor covalent bond with bond order 1 by the formation of an H-bond with acceptor unit Y in which electron density is transferred from the electronegative, electron rich, Y unit into a σ^* anti-bonding X–H orbital. The effect is a weakening of the X–H link, an increase in the X–H distance and a polarisation of the bond, enabling the formation of a link with Y via the intermediate H atom.

This explains the broad features of the so-called red-shift in infra red (IR) spectra characteristic of H-bonding, which provided the first source of direct insight into the character of H-bonds at the microlevel. Microtheories must be consistent with the thermodynamic data, which is neutral with respect to microtheory and gives no direct handle on the character of H-bonds. The first investigations of IR spectra of H-bonding in the 1930s were taken to show that the characteristic vibrational stretching frequency of the X–H covalent donor bond disappears on formation of an X–H•••Y bond. But shortly afterwards it was realised that it is red-shifted to a lower frequency. The weakening and lengthening of the X–H bond account for the red-shift of the vibrational stretching frequency, v_s(X–H), to a lower wave number and the increased polarisation accounts for the increased intensity of the IR absorption, which depends on the change in dipole moment during vibration. A more specific account of the detailed features of the spectrum calls for a more detailed theoretical model, which cannot be pursued here.

The calculation of bond energies is an important application of theoretical models. For an H-bonded dimer, the bond energy is defined as the difference between the energy of the bound dimer and sum of the energies of the separate monomers. This difference is usually small in relation to the energies of the separate monomers, and so calls for much refinement of the approximation procedures employed in order that the error in the monomer energies shouldn't magnify into too great an error in the bond energy. A much-used approach analyses the bond energy in terms of the Morokuma decomposition. Although decomposing into components – a standard procedure on the VB approach – is somewhat artificial in MO theory, Morokuma (1971) developed such an analysis using Hartree-Fock calculations in which bond energy, ΔE, is given by

$$\Delta \mathrm{E} = \mathrm{S} + \mathrm{EX} + \mathrm{PL} + \mathrm{CT} + \mathrm{MIX}.$$

Here ES is the purely electrostatic stabilising interaction between two unmodified (frozen) monomer charge densities. It is countered by EX, the exchange repulsion due to the mixing of the occupied orbitals of both monomers and the corresponding action of the Pauli exclusion principle. It removes electron density from the interaction zone, preventing two monomers from approaching each other too closely. PL is the polarisation, a stabilising effect due to the approach of one molecule modifying the electronic structure of X–H in the other molecule (by mixing the occupied and the vacant orbitals of X–H) at the same time as the approach of X–H modifies the electron density distribution over the first molecule. CT is the charge transfer arising from mixing of vacant orbitals of one partner with occupied orbitals of the other, resulting in a net electron transfer from proton acceptor to proton donor. Because these factors are not independent, a coupling factor MIX is added, which is an increasing function of ΔE.

On the basis of this analysis, Vanquickenborne (1991) has compared the energies of H-bonds linking dimers falling into two groups, one with bond energies in the weak to moderate range (4.1 kcal mol^{-1} for $H_2NH{\bullet}{\bullet}{\bullet}OH_2$ to 16.3 kcal mol^{-1} for $FH{\bullet}{\bullet}{\bullet}NH_3$), and the other with bond energies strong enough to affect the geometry of the monomers (27.3 kcal mol^{-1} for $H_2O{\bullet}{\bullet}{\bullet}NH_4{}^+$, and 62.7 kcal mol^{-1} for $H{\bullet}{\bullet}{\bullet}HF^-$). The contributing factors follow substantially different patterns in the two groups, but even within the groups the trends are not simple when it comes to details. In the first group, all energy components are increasing functions of ΔE, suggesting that essentially the same type of bonding is at issue in all cases, although ΔE cannot be construed as a linear function of any of the Morokuma components. In the second group, the charge transfer contribution is unusually large in $HF_2{}^-$, accounting for 21.4 % of the attractive energy (44 % of ΔE) and exceeding the difference between the attractive ES and repulsive EX contributions. The exchange destabilisation is unusually small in $H_2O{\bullet}{\bullet}{\bullet}NH_4{}^+$. It is results of this kind that underlie the IUPAC Task Group comment that H-bonding cannot be attributed to a "single physical force".

Another way of looking at bonding is to consider the electron density distribution. Just as the binding energy is the difference between the energies of the dimer less that of the separate monomers, so we can see how the electron density shifts by considering distributions before and after bond formation, and calculating the dimer density minus the sum of the two monomer densities as a continuous function, $\Delta\rho$. Maps of $\Delta\rho$ for the water dimer show a clear depopulation around the bridging H atom, making the positive H atom more positive upon bridging. This nicely illustrates the special character of H-bonding when compared with density difference maps of standard covalent and ionic bonds.

The interaction in a typical covalent bond is marked by an accumulation of electron density as electrons flow into the central bonding zone. A similar density difference map of sodium fluoride compared to the isolated atoms shows how electron density is transferred from the sodium atom to the fluorine atom, as we would expect for an ionic bond. But where the comparison in the difference map is between the bonded sodium fluoride ions and the separate *ions*, rather than the isolated *atoms* – i.e. after the electron has been transferred – the electron density is seen to flow back into the central bonding zone, illustrating that even in NaF there is a certain amount of covalent character. The depletion of electron density around the H-bridge of an H-bond resembles the relation between the non-bonding pair of noble gas atoms rather than the standard covalent and ionic bonds with their accumulations of electron density between the bonding atoms. Several authors point to this as a feature distinguishing H- from typical bonds, be they covalent or ionic.

5 Concluding Remarks

Drawing a dividing line in the range of phenomena displaying chemical affinity between bonding and nonbonding interactions is not as clear-cut as it might once have seemed. Nevertheless, it has been argued that the directional character of H-bonding, together with its occurrence as an intramolecular force and exhibiting a range of bond strengths extending substantially beyond the weak bonds originally thought characteristic of H-bonding, suffice to count H-bonds as bonds. A corresponding account of the forms of covalent bonding in compounds deviating from Lewis's scheme – itself a marked departure from the simple representation of bonds as sticks connecting atomic symbols – would add weight to this conclusion.

Remembering Coulson's scepticism might suggest that the conclusion be tempered in some way. But formulating the issue as a question of existence, as Coulson does, is unfortunate since the existence of bonds might be denied on more general grounds having nothing to do with the specific chemical issues about bonding. What exists are entities like electrons and nuclei, of which it might be true that they sustain properties and relations such as being bonded. Bonding might be construed as a process which could be said to exist. But I take it this is not the case. There are processes in which specific features of the bonding play a crucial role, such as chemical reactions like those involving the transfer of hydrogen ions and atoms in an aqueous medium, but bonding is not itself such a process. So unless bonding is to be clearly identified with a body such as a pair (or mereological sum) of electrons, the question of the existence of bonds doesn't arise and the issue is rather whether a certain condition is satisfied.

Taking a bond to be a body such as a pair of electrons raises difficulties because the quantum nature of identical particles involved suggests that it won't be possible to say which body is the bond. Weisberg takes up this theme, suggesting that it is not even possible to understand bonds more generally as "submolecular regions of electron density" because of "delocalization – density spread beyond the submolecular region between the atoms ..." (2008, pp. 944, 945). In the case of H-bonding, the density difference maps exacerbate this problem. Some sort of build up of electron density between the bonded atoms is usually to be found despite the delocalisation. But density difference maps suggest that electron density diminishes around the hydrogen when it participates in H-bonding.

Hendry speaks of Coulson's sceptical thoughts about the existence of bonds leading him to focus on changes in energy rather than "seeking a material element that realizes the theoretical role of keeping a molecule together" (2008, pp. 918–919). This would be to construe bonding as a relation which obtains between molecular subunits or as a property of the molecule as a whole provided a condition to the effect that the energy of the molecule is substantially less than that of the isolated subunits is satisfied. The concept of bond energy can be problematic in the case of H-bonding, however. We saw how it is defined when linking monomers in dimers as the difference between energies of isolated and combined (dimer) states. This works for *inter*molecular H-bonds. But in the case of *intra*molecular

H-bonding, the bond energy is not defined because it is not possible to define energy as the difference of two appropriate states.

Bringing H-bonding into the picture makes the general conception of a bond even more elusive than Weisberg and Hendry already suggest it is. We have seen how Pauling's definition of H-bonding came to be seen as too narrow, why Pimentel and McClellan's characterisation was criticised for encompassing too much, and how Arunan et al. (2010) have tried to more precisely ring in the appropriate phenomena with an open-ended formulation that strives to accommodate the theoretical and experimental insights on which criticisms of earlier proposals were based. The Task Group's proposal is not a definition in the formal sense, and doesn't even point in the direction of a disjunctive definition of "higher-level" chemical properties in terms of "lower-level" properties featuring in theories of physics. Although Sober (1999) shows that disjunctive complexity is no obstacle to reduction, this doesn't settle whether bonding is just physics "but merely pushes the question back one step" (Sober 1999, p. 562). As matters stand, the Task Group's proposal integrates resources from physics and chemistry, with some pointers to explanatory depth, others to explanatory generality (in Sober's sense), but without firm indications of reductionism. This is the unity displayed by extant science, and I see no interest in idle speculation about whether H-bonding, bonding in general, let alone chemistry as a whole, will collapse into physics proper by elimination or reduction in the dim and distant future.

References

Arunan, E., Desiraju, G., Klein, R., Sadlej, J., Scheiner, S., Alkorta, I., Clary, D., Crabtree, R., Dannenberg, J., Hobza, P., Kjaergaard, H., Legon, A., Mennucci, B., & Nesbitt, D. (2010). *Defining the hydrogen bond: An account*. IUPAC Technical Report. http://media.iupac.org/reports/provisional/abstract11/arunan_tr.pdf. Accessed 12 July 2011.

Coulson, C. A. (1955). The contributions of wave mechanics to chemistry. *Journal of the Chemical Society,* 2069–2084.

Coulson, C. A. (1961). *Valence*. London: Oxford University Press.

Hendry, R. (2008). Two conceptions of the chemical bond. *Philosophy of Science, 75*, 909–920.

Latimer, W., & Rodebush, W. (1920). Polarity and ionization from the standpoint of the Lewis theory of valence. *Journal of the American Chemical Society, 42*, 1419–1433.

Morokuma, K. (1971). Molecular orbital studies of hydrogen bonds, III. C=O–H···O hydrogen bond in $H_2CO\cdots H_2O$ and $H_2CO\cdots 2H_2O$. *Journal of Chemical Physics, 55*, 1236–1244.

Pauling, L. (1960). *The nature of the chemical bond and the structure of molecules and crystals* (3rd ed.). New York: Cornell University Press.

Pimentel, G., & McClellan, A. (1960). *The hydrogen bond*. San Francisco: Freeman.

Sober, E. (1999). The multiple realizability argument against reductionism. *Philosophy of Science, 66*, 542–564.

Sutcliffe, B. (1996). The development of the idea of a chemical bond. *International Journal of Quantum Chemistry, 58*, 645–655.

Vanquickenborne, L. (1991). Quantum chemistry of the hydrogen bond. In P. L. Huyskens, W. A. P. Luck, & T. Zeegers-Huyskens (Eds.), *Intermolecular forces: An introduction to modern methods and results* (pp. 31–53). Heidelberg: Springer.

Weisberg, M. (2008). Challenges to the structural conception of bonding. *Philosophy of Science, 75*, 932–946.

The Metaphysics of Molecular Structure

Robin Findlay Hendry

Abstract In this paper I distinguish two kinds of structure that appear in chemical explanations: geometrical structure and bond structure. I examine structural descriptions of a range of chemical substances including sodium chloride, water, cyclohexane and proteins, arguing that neither kind of structure is more basic than the other. Such pluralism should not be surprising, however, because at least in chemistry, structure is a creature of classification, and therefore of comparison.

1 Introduction

Chemists appeal to structure at the molecular level—molecular structure—to explain the thermodynamic, chemical and spectroscopic behaviour of chemical substances. Structure is the sole basis of the systematic nomenclature by which substances are named (Thurlow 1998). But what is a molecular structure? In general, the structure of a thing is how its parts fit together. The parts of molecules are atoms and ions, which leaves how they fit together. In fact chemical explanations invoke two kinds of structure, which I will call geometrical structure (the relative positions of the atoms and ions) and bond structure (the framework of bonds between the atoms and ions). The two kinds of structure are perfectly reconcilable, and some substances have both. But they are quite distinct. In what follows I will describe them, and the relationships they bear to each other. In the final section I will argue for pluralism about structure, and that this should not be surprising, given that structure is primarily a classificatory notion.

R.F. Hendry (✉)
Department of Philosophy, Durham University, 50 Old Elvet, Durham, DH1 3HN UK
e-mail: r.f.hendry@durham.ac.uk

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_27,

2 Geometrical Structure

When chemists describe the ‘structure’ of a substance, at least sometimes they mean something that can be specified fully in terms of the (average) relative positions of the constituent atoms and ions. Sodium chloride (NaCl) is—pretty much—positively-charged charged sodium ions and negatively-charged chloride ions in a one-to-one ratio. *Solid* NaCl is composed of ‘two interpenetrating face-centred cubic sub-lattices’ (Greenwood 1968, p. 48), in each of which each a sodium (or chloride) ion is surrounded by six chloride (or sodium) ions arranged octahedrally. So it may be considered as a (potentially infinite) array of unit cells, each cell containing four sodium ions and four chloride ions (see Fig. 1).

There are four of each kind of ion in a cell because the eight ions at the corners are each shared with seven other unit cells, so each counts only as one eighth; the 12 ions at the edges are each shared with three other cells, so each counts as one quarter; the six ions at the faces of the cube are shared with one other cell, so each counts as one half; and finally the ion at the centre falls entirely within the cell, so counts as 1.

The structure of an ionic solid arises from the way the constituent ions pack together so as to maximise interactions between ions of opposite charge, and minimise interactions between those of like charge, given the charges on the ions, the relative size of the cations and anions (M^+ and X^-, respectively), and the stoichiometry of the substance (i.e. whether it is of the form M_2X, MX, MX_2 or so on). Although the structure is characterised by the relative positions of the ions, as represented by distances between the ionic centres (which can be regarded as the sum of two ‘ionic radii’), the ions are not entirely static: they vibrate around their equilibrium positions to an extent that is dependent on temperature, so the distances fluctuate. Since the structure survives such fluctuations, it must be characterised by small regions around *average* relative positions. At 801 °C, however, enough of the ions have enough energy to overcome the forces holding them in the lattice and the structure breaks down, forming a liquid consisting mostly of dissociated ions: since the ions are now free to move under electrical forces, the molten salt is an electrical conductor while the solid is an insulator. Clearly, the geometrical structure of solid

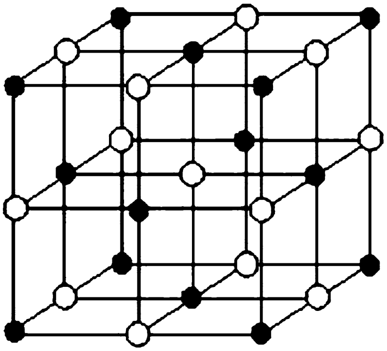


Fig. 1 Solid sodium chloride (After Greenwood 1968, p. 48)

NaCl does not survive transition to the liquid phase: it is phase-specific. Molten NaCl, like other liquids, has its own structure, which can be characterised in terms of radial distribution functions describing the probability density of various molecular or atomic species as a function of their distance from a central atom. Once again, the structure is fully specified by geometrical relations between the constituent ions, and is phase-specific, in that it exists only within a particular state of aggregation. Water is similar. Depending on pressure, ice is described as displaying one of a number of different structures (see Eisenberg and Kauzmann 1969, Chap. 3; Finney 2004), in all of which hydrogen bonds play an essential role (Needham 2013), linking together the partial negative charges on oxygen atoms to the partial positive charges of protons on neighbouring H_2O molecules. As in NaCl, this structure breaks down on transition to the liquid phase. It is not that the H_2O molecules cease to form hydrogen bonds with each other, or that these bonds cease to constrain their relative positions and orientations: it is rather that, in this higher temperature range, the H_2O molecules are freer to move around them, and the hydrogen bonds themselves are constantly forming and reforming. So even though, at short range, the structure of liquid water is quite like that of ice, over longer ranges this breaks down, as displayed in the radial distribution functions used to describe its structure (see Fig. 2).

If a structure is constituted by the *average* relative positions of the atoms or ions, then structure in this sense must depend on the energy range and timescale over which that average is taken. The cell structure of solid NaCl, as we saw, breaks down above its melting point, and so if we choose a wide enough energy range, the long-range geometrical order of solid NaCl is lost. Similarly, once it is acknowledged that even in the solid state, atoms and ions are constantly in motion, ‘structure’ depends on timescale. Eisenberg and Kauzmann (1969, 150–152) point out that H_2O molecules in ice undergo vibrational, rotational and translational motions, the molecules vibrating much faster than they rotate or move through the lattice. At very short timescales (shorter than the period of vibration), the ‘structure’ is a snapshot of molecules caught in mid-vibration. It will be disordered because different molecules will be caught at slightly different stages of the vibration. As timescales get longer, the ‘structure’ averages over the vibrational motions, and then the rotational and translational motions, yielding successively more regular but diffuse structures. None of this should be surprising: different kinds of structural feature persist over different energy ranges and timescales, and so the energy ranges and timescales we focus on in some particular case will determine which structural features are part of ‘the’ structure (alongside, of course their relevance to the things we want to explain).

3 Bond Structure

The bond structure of a substance, the framework of bonds between its constituent atoms or ions, is quite different from its geometrical structure, which is constituted by geometrical relationships between them. Consider, for instance, cyclohexane, which is a cyclic alkane—a hydrocarbon involving only single bonds between

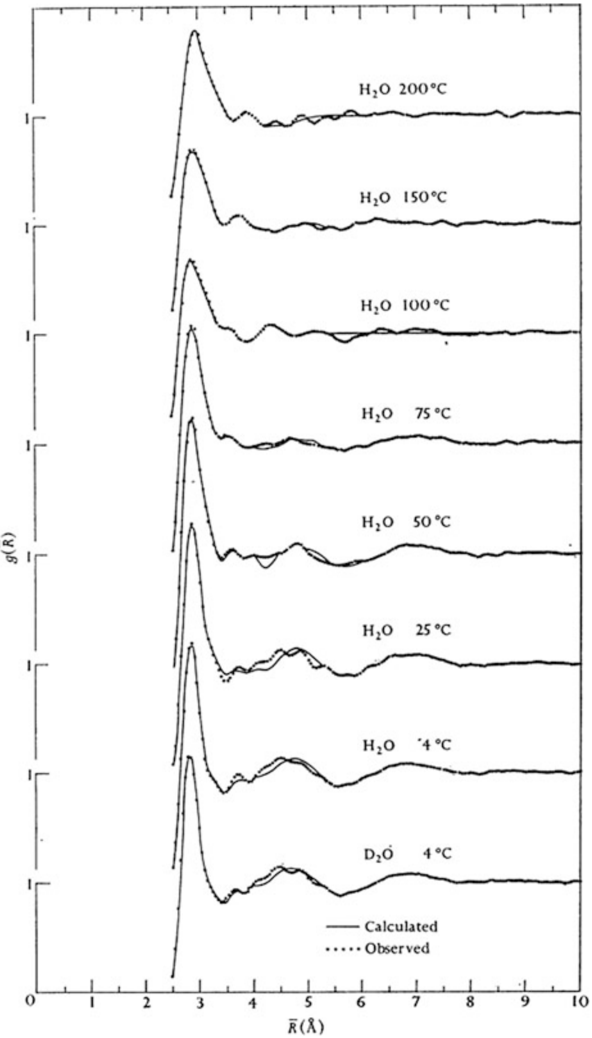


Fig. 2 Radial distribution functions at various temperatures for H_2O and D_2O, (Reproduced from Eisenberg and Kauzmann 1969, p. 157)

carbon atoms in a ring structure—with empirical formula C_6H_{12}. In cyclohexane the six carbon atoms are bonded together in a ring, and to each is attached two hydrogen atoms (see Fig. 3).

The bond structure of cyclohexane is easily distinguished from its geometrical structure. Firstly, consider any pair of hydrogen atoms which are attached to the same carbon atom. These two hydrogen atoms may be geometrically adjacent to each other in the sense that they are not far apart, and no other atom is between them (they are in each other's line of sight). But they are not bonded directly to each other. Secondly, the bond structure is compatible with wide variation in the relative positions of the atoms, and different geometrical structures. Cyclohexane

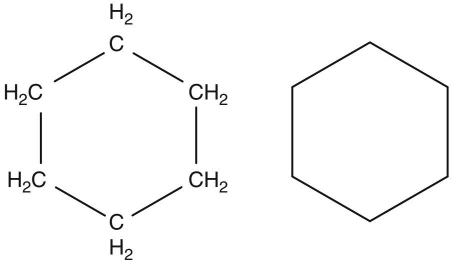
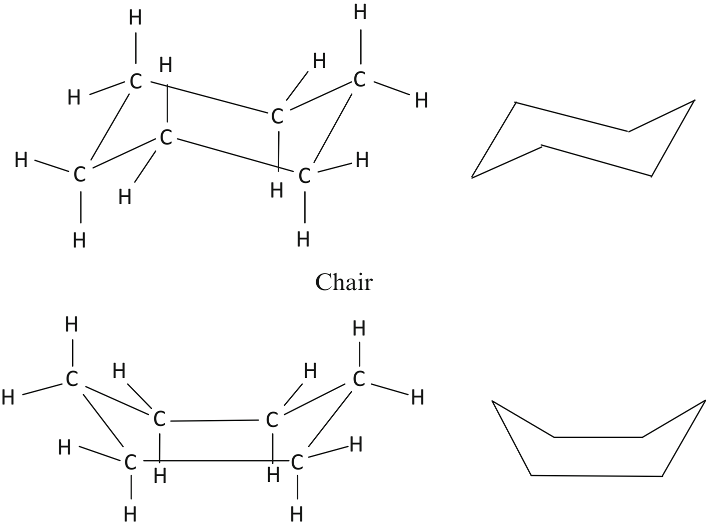


Fig. 3 Two representations of the bond structure of cyclohexane: in the image on the right, the hydrogen atoms are left out for clarity



Chair

Boat

Fig. 4 Conformations of cyclohexane: in the images on the right, the hydrogen atoms are left out for clarity

exhibits a number of different *conformations*: that is, geometrical configurations of its atoms. Cyclohexane's lowest energy conformation is the chair (see Fig. 4), but individual cyclohexane molecules are constantly in motion. The energy difference between chair and boat is small, and molecules flip between them many thousands of times a second.

Across all the different conformations, however, one thing remains constant: the pattern of connections between the atoms. This is the bond structure. In the 1860s there appeared a number of different but equivalent ways of representing the bond structure of molecules (see Rocke 1984, 2010), which employed either diagrams on paper or three-dimensional models. They were equivalent in the sense that the structures they represented, attributed on the basis of chemical evidence, were topologically identical. They were constructed under rules of valence which determined, for each element, how many atoms of the various other types it could be bonded to in a molecule. The topological nature of these bond structures was recognised explicitly in Arthur Cayley's discussion of these 'chemical graphs' as 'trees', his application of this to isomerism, and his formal proof of how many distinct aliphatic hydrocarbons there are with empirical formula C_nH_{n+2} (see Biggs et al. 1976, Chap. 4).

By the mid-1870s, graphical formulae came to be understood as embedded in three-dimensional space. The embedding made available new kinds of chemical evidence for distinguishing between structures. Jacobus van't Hoff explained why there are two isomers of compounds in which four different groups are attached to a single carbon atom by supposing that the valences are arranged tetrahedrally (the two isomers are conceived of as mirror images of each other). Adolf von Baeyer explained the instability and reactivity of some organic compounds by reference to strain in their molecules (Ramberg 2003, Chaps. 3 and 4), which meant distortion away from their preferred geometry. These stereochemical theories were intrinsically spatial, because their explanatory power depended precisely on their describing the arrangement of atoms in space. From the beginning of the twentieth century, bond structures became dynamic, as chemists and physicists began to develop models of how molecules vibrate and rotate, to explain their spectroscopic behaviour (Assmus 1992). This involved filling out structures with details, such as bond lengths, bond angles and force constants, which had previously been absent.

The valence rules have a curious status. They provided a reliable guide to the development of organic chemistry, successfully attributing structures to a vast number of chemical substances, and many of the structures attributed to substances in the 1860s are still accepted in modern chemistry. G.N. Lewis thought 'this group of ideas which we call structural theory' (1923, 20–21) to be one of the most successful in science, yet he recognised that chemical substances did not always behave (physically or chemically) in accordance with their structural formulae: this was a problem especially in inorganic chemistry (1923, p. 67). Moreover, there were always well-known anomalies: substances, like carbon monoxide, in which some atom does not display its usual valence. Whether they cover all of chemistry, or just some well-behaved fragment of it, the idea of these valence rules is worth exploring: they assign valences to atoms according to their elemental identity, and require that, in a valence structure, all of an atom's valences should be used up in single or multiple connections to other atoms. They govern the construction of graphs, and so they should be expressible in first-order logic: there must be an axiomatisation of what chemists call the 'classical theory of molecular structure,' even though that theory remained entirely implicit in the nineteenth century.[1]

[1] Note that the theory is not 'classical' in any way that connects it with 'classical mechanics.' In particular, it is not a dynamical theory, and so is not governed by Newtonian dynamics; it is not a statistical theory, and so does not assume that the atoms over which it quantifies are governed by Boltzmann (or any other) statistics.

4 Geometrical Structure Versus Bond Structure

Clearly, geometrical structure and bond structure are not the same thing. What is the relationship between them? Is one more basic or fundamental than the other?

Some substances may have a geometrical structure yet (arguably) lack a bond structure. G.N. Lewis established this when he argued that the structure of ionic substances like potassium chloride (KCl) can be represented without appeal to any bonds between atoms. Lewis (1913) considers a proposal to represent ionic bonding in potassium chloride with a directed arrow, as K→Cl, which would signify that an electron has passed from K to Cl. He argues that this would be misleading, because *even if* (*per impossibile*, given the qualitative identity of electrons), one could tell which electron had come from which potassium atom, the bonding that holds the substance together does not arise from that donation, but rather from the opposite charges that result from it. Furthermore, 'a positive charge does not attract one negative charge only, but all the negative charges in its neighborhood' (Lewis 1913, p. 1452). In potassium chloride, the bonding is electrostatic and therefore radially symmetrical. An individual ion bears no *special* relationship to any *one* of its neighbours, but the same relationship to each of them. This relationship is non-directional, and so cannot be represented by the lines connecting atoms that appear in classical structural formulae. Nor did Greenwood's description of NaCl mention bonds (see above): bonds are not indispensible to a description of its structure. So even though, in the representation of the structure of NaCl (Fig. 1, above), we can see lines between neighbouring ions, they are merely an aid to the eye in discerning the three-dimensional structure of a unit cell. The lines do not represent real physical features of NaCl's structure. If this is right, then some substances have a geometrical structure but no bonds, and therefore no bond structure. How is it possible to have bonding without bonds? There is certainly bonding, because the ions are held together in the lattice by something or other (to a large extent, electrostatic attraction). So although here is a 'bond' in one abstract sense of the word (as in the phrase 'the bond of sisterhood'), there is no bond in the sense which is important to Lewis' argument: the pairwise physical relationship between individual atoms or ions, which is represented by the lines between atoms in molecular structure diagrams. Lewis' argument is meant to establish that there is geometrical structure in NaCl, but no bonds in that second sense of 'bond,' and therefore no bond structure.

Furthermore, every molecule has a geometrical structure, in the sense that its parts are distributed somehow in space, and they bear spatial relations to each other. Given that not every substance has a bond structure, this seems to favour geometrical structure over bond structure for the leading role in the relationship between them: geometrical structure is a more general, and so more basic notion, because having a geometrical structure is necessary, even if not sufficient, for having a bond structure. But that would be misleading for two reasons. Firstly, it is not so clear that having a geometrical structure is necessary for having a bond structure, at least in any way that would make it more basic. From a mathematical point of view, a bond structure is a set-theoretic object: if we take the set of a molecule's constituent atoms, a bond structure is some subset of the set of ordered pairs that can be formed from the members of this set. This set-theoretic structure is all that is needed to fulfil one important explanatory role for bond structure in chemistry: that of explaining how many structural isomers a particular substance may have. And from a purely logical point of view, something might have this set-theoretic structure without it (or its parts) being located in space at all. This is just how the explanatory role of structure was seen by the pioneers of structure in chemistry in the 1860s, such as Edward Frankland (see Hendry 2008b). Even though bond structures did eventually come to be regarded as embedded in space, that was an extension of the explanatory role of structure to account for optical isomerism. From a purely mathematical point of view, then, geometrical relationships do not determine bonding relationships. Perhaps bond structure is only contingently embedded in space. But the mathematical point of view is not all there is, and a bond structure is not just a graph: it is a graph generated by a particular *physical* relation (the bonding relation). Is a *bond* structure something that is necessarily embedded in space? To answer that question we need to know more about what a bond is. (That is not merely a rhetorical pointer to a later discussion: the answer is not clear: in Hendry 2008b, 2010 I discuss two opposed accounts.) Bonds clearly have geometrical constraints: distinct bonds do not overlap or cross, and it may well be that fixing the geometrical configuration *physically* (though not mathematically) determines the bond structure uniquely, in the following way.

In the ‘Atoms in Molecules’ (AIM) programme, Richard Bader and his co-workers have sought to recover the traditional bond structure of molecules as a topological feature of the electron-density distribution (see Bader 1990; Popelier 2000; Gillespie and Popelier 2001). From the electron-density distributions for many different molecules can be defined ‘bond paths’ between atoms that generate ‘molecular graphs’ which are strikingly close to the classical molecular structures of those molecules. As Bader puts it, ‘The recovery of a chemical structure in terms of a property of the system’s charge density is a most remarkable and important result’ (1990, p. 33). Bader’s elegant results are interesting and significant in a number of ways. Firstly, AIM offers a substantive answer to a longstanding question: what is a chemical bond? The answer (according to AIM) is that bonds are topological features of the electron density distribution (or rather the particular regions of electronic charge that bear these topological features). Secondly, although it recovers the classical bond topology, AIM seems to make geometrical structure prior to bond structure. The quantum-mechanical calculations that underlie AIM, like all tractable quantum-mechanical calculations concerning molecules, begin by making the Born-Oppenheimer approximation (see Hendry 1998, forthcoming), which involves separating nuclear and electronic variables, and fixing (or ‘clamping’) the nuclear positions. The electric field due to the nuclei is then used as a constraint on the calculation of a resultant electron density distribution. If the nuclear positions are well chosen (i.e. correspond to the nuclear positions in the molecule’s equilibrium geometry), then from the resulting electron density distribution we have ‘read off’ the bond structure of a real molecule from its nuclear geometry. Physically, if not mathematically, it might seem that geometry determines bond structure. But that is too quick: it's not so clear that we can simply 'read off' the bond structure from the geometry, because the whole calculation relies on minimising the energy of the system. (That, in fact, is taken to be a mark of how closely the Born-Oppenheimer calculation approximates the 'exact' energy.) By concentrating on the lowest-energy states, the whole procedure would seem simply to ignore higher-energy states that correspond to higher-energy geometrical configurations of different bond topologies. Perhaps we don't find *the* unique bond topology for the geometry, but rather the bond topology which has the lowest energy in that geometry (and, probably, has that geometry as its lowest-energy geometrical configuration).

Let us pursue this idea that a bond structure is something that can be displayed by a substance *in addition* to its geometrical structure. This is supported by the fact that bond structure may survive phase transitions which the geometrical structure cannot. Thus, for instance ice, liquid water and steam all display different geometrical structures, but the topological structure of its molecules, as represented in its structural formula (a central oxygen atom bonded to two hydrogen atoms) remains constant across the different states of aggregation. Secondly, in the substances that have it, bond structure is explanatorily prior, in the sense that a molecule's bond structure is compatible with a range of different geometrical arrangements of its parts, and in fact determines which arrangements it may have. Consider once again the conformations of cyclohexane. In that case, the bond structure is a constant while the molecule moves between quite different geometrical configurations. And it is the persistent bond structure which *explains* the energetic ordering of the various conformations. The chair is the lowest-energy conformation because in that geometry the bond structure experiences the least strain: that is, in that geometry the arrangement of bonds around individual carbon atoms is closest to the tetrahedral, and the hydrogen atoms are less crowded, reducing their (repulsive) interactions. These considerations allow us, I think, to resist the idea that geometrical structure is prior to bond structure.

5 Structure as Abstraction

If neither geometrical structure nor bond structure is prior to the other, how might one understand the role of these two notions in the classification of substances, and in explanations of their behaviour and characteristic properties as arising from their structure? Seen as a classificatory notion, structure is derived from a process of abstraction. Molecules that differ in the properties of, and relations between, their parts can share a structure. Identifying a class of molecules as sharing structure, we just ignore the differences between them. Different classes of molecule or substance may be alike in different ways: we might expect there to be different kinds and levels of structure. We get to the structure of a substance by abstracting away from the particular clusters of property- and relation-instances that its constituent atoms and ions bear to each other, to focus on some subset of them which is salient because it survives across some range of (e.g. thermodynamic or energetic) conditions that demands our attention. Pluralism about structure should not then be surprising simply because in a reasonably complex thing there is always, in principle, more than one way to abstract away from its full particularity in this way.

Take water and proteins as examples.[2] Water, as we have seen, has both a bond structure and a geometrical structure, although the bearers of these structures are different (individual molecules as opposed to collections of such molecules). Considered as a substance which can exist in different states of aggregation, we focus on the covalent bond structure of its molecules that is shared by all those different states (solid, liquid, and gas, up to highly rarefied states). You cannot abstract away from that bond structure without abstracting away from water itself (or so I argue: see Hendry 2006, 2008a). But hydrogen bonding is 'structural' too: in water, it plays an important role in understanding the structure of the substance in its particular states of aggregation. Neither is *the* structure, in the sense that, if we need to know about the details of water's structure in one of its particular states of aggregation, the bond structure won't tell us enough. It is the interaction between the individual molecules that gives rise to the (large-scale) geometrical structure, via hydrogen bonding (on which see Needham 2013): the opposite partial charges on oxygen and hydrogen atoms give rise to chains and clusters of H_2O molecules whose existence has a major influence on the properties of the substance. But these chains and clusters are constantly forming and reforming.

Proteins also display different kinds and levels of structure. The primary structure of a protein is, roughly, the order of the connections between its constituent amino-acids, while secondary, tertiary and quaternary structure concern the different ways in which it is arranged in space. Interestingly, the very same physical interaction, hydrogen bonding, which gives rise to short-lived structure in water, maintains structure that is longer-lasting in proteins, at least in the narrow range of physical and chemical conditions within which cellular processes take place. But these higher levels of structure do not survive at higher temperatures, or in more hostile chemical environments. Concentrating on temperature, one might say that the higher levels of structure are conformations 'frozen in' below the temperature at which the hydrogen bonds that sustain them would begin to break and reform too quickly. Important phenomena (biology!) depend on them, but if you consider a wide enough range of conditions, the higher levels of structure disappear from view.

I conclude that, in identifying different kinds of structure and reasoning about them, we (implicitly) focus on relations which survive over the specific ranges of (chemical and physical) conditions in which the phenomena of interest can be given a unified explanation in terms of that kind of structure. Since there is a close relationship between structure and substance identity, these specific ranges of conditions are essentially those at which identifiable substances exist. Different substances are stable over different ranges of physical conditions, and

[2] See Tobin (2010) and Slater (2009) for discussion of classification and structure in proteins; Goodwin (2011) provides a reply.

it should be no surprise if structural explanations concerning substance X focus on physical interactions that underlie the particular structural relationships that survive across the conditions under which X exists, and structural explanations concerning substance Y focus on different physical interactions that underlie the different structural relationships that survive across the different conditions under which Y exists.

6 Conclusion

'Structure' sometimes invokes geometrical relations. Sometimes it invokes bond topology, which is understood always to be embedded in space. Furthermore neither kind of structure is more basic than the other. If these arguments support some form of pluralism, it is one that should take a robustly realist stance on structure and its role in classification. It is robustly realist for two reasons. Firstly, each kind of structure is constituted by real physical relations: spatial relations in the case of geometrical structure, bonds in the case of bond structure. Secondly, you can't ignore either kind of structure without significant loss of information about the substances that have them. If you ignore geometrical structure, you have no access to the explanations provided, for instance, by optical isomerism (van't Hoff), or stearic strain (von Beayer). If you ignore bond structure you ignore what, in many substances, is held constant over a range of different geometrical configurations (remember the conformations of cyclohexane), and explains which geometries are possible, and which are favoured energetically.

References

Assmus, A. (1992). The molecular tradition in early quantum theory. *Historical Studies in the Physical and Biological Sciences, 22*, 209–231.

Bader, R. (1990). *Atoms in molecules: A quantum theory*. Oxford: Oxford University Press.

Biggs, N. L., Lloyd, E. K., & Wilson, R. J. (1976). *Graph theory 1736–1936*. Oxford: Clarendon Press.

Eisenberg, D., & Kauzmann, W. (1969). *The structure and properties of water*. Oxford: Oxford University Press.

Finney, J. L. (2004). Water? What's so special about it? *Philosophical Transactions of the Royal Society B, 359*, 1145–1165.

Gillespie, R., & Popelier, P. (2001). *Chemical bonding and molecular geometry*. Oxford: Oxford University Press.

Goodwin, W. (2011). Structure, function, and protein taxonomy. *Biology and Philosophy, 26*, 533–545.

Greenwood, N. N. (1968). *Ionic crystals, lattice defects and nonstoichiometry*. London: Butterworths.

Hendry, R. F. (1998). Models and approximations in quantum chemistry. In N. Shanks (Ed.), *Idealization in contemporary physics* (pp. 123–142). Amsterdam: Rodopi.

Hendry, R. F. (2006). Elements, compounds and other chemical kinds. *Philosophy of Science, 73*, 864–875.

Hendry, R. F. (2008a). Microstructuralism: Problems and prospects. In K. Ruthenberg & J. van Brakel (Eds.), *Stuff: The nature of chemical substances* (pp. 107–120). Würzburg: Königshausen und von Neumann.

Hendry, R. F. (2008b). Two conceptions of the chemical bond. *Philosophy of Science, 75*, 909–920.

Hendry, R. F. (2010). The chemical bond: Structure, energy and explanation. In M. Dorato, M. Rèdei, & M. Suárez (Eds.), *EPSA Philosophical issues in the sciences: Launch of the European Philosophy of Science Association* (pp. 117–127). Berlin: Springer.

Hendry, R. F. (forthcoming). *The metaphysics of chemistry*. Oxford: Oxford University Press.

Lewis, G. N. (1913). Valence and tautomerism. *Journal of the American Chemical Society, 35*, 1448–1455.

Lewis, G. N. (1923). *Valence and the structure of atoms and molecules*. Washington: Chemical Catalogue Company.

Needham, P. (2013). Hydrogen bonding: Homing in on a tricky concept. *Studies in History and Philosophy of Science, 44*, 51–66.

Popelier, P. (2000). *Atoms in molecules: An introduction*. London: Pearson.

Ramberg, P. (2003). *Chemical structure, spatial arrangement: The early history of stereochemistry 1874–1914*. Aldershot: Ashgate.

Rocke, A. (1984). *Chemical atomism in the nineteenth century: From Dalton to Cannizzaro*. Columbus: Ohio State University Press.

Rocke, A. (2010). *Image and reality: Kekule, Kopp and the scientific imagination*. Chicago: Chicago University Press.

Slater, M. (2009). Macromolecular pluralism. *Philosophy of Science, 76*, 851–863.

Thurlow, K. J. (1998). IUPAC Nomenclature part 1, organic. In K. J. Thurlow (Ed.), *Chemical nomenclature* (pp. 103–126). Dordrecht: Kluwer.

Tobin, E. (2010). Microstructuralism and macromolecules: The case of moonlighting proteins. *Foundations of Chemistry, 12*, 41–54.

Part IX
Philosophy of the Life Sciences

Description, Explanation, and Explanatory Depth in Developmental Biology

Christopher H. Pearson

Abstract The last few decades have seen molecular genetics occupy an expanding role in developmental biology. Alexander Rosenberg has argued that developmental biology's shift to articulating the molecular basis for organismic development represents the point at which developmental biology becomes an explanatory discipline. This essay is a critical response to Rosenberg's view, one that works to show that developmental biology is rich with explanatory resources in the absence of molecular genetics. At the same time, the essay seeks to articulate, by way of an appeal to explanatory depth, the explanatory value molecular genetics often provides to developmental biology.

1 Introduction

There is little question that the focus of working developmental biologists has trended strongly towards attention to the underlying genetics operating in the development of organisms. Alexander Rosenberg (1997, 2006) has, in turn, offered an admirably detailed account of the philosophical import of that trend in his attempt to demonstrate how developmental biology's shift in focus to molecular genetics supports a reductionistic view of developmental biology. In the course of outlining the reasons for accepting reductionism, Rosenberg also advances a rather provocative thesis about the explanatory character of developmental biology. Specifically, Rosenberg contends that, in addition to molecular genetics ushering in a reduction of developmental biology to molecular genetics, the integration of molecular genetics to developmental biology radically transforms developmental

C.H. Pearson (✉)
Department of Philosophy, Southern Illinois University Edwardsville, Box 1433,
Edwardsville, IL 62026, USA
e-mail: chpears@siue.edu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_28,

biology with respect to its ability to explain. At various points in Rosenberg's writing, the transformation that he sees molecular genetics initiating in developmental biology is described somewhat differently. A strong interpretation of Rosenberg's view, however, might characterize the difference between pre and post molecular (genetics) developmental biology as one that centers on descriptive versus explanatory practices.[1]

The distinction between descriptive and explanatory practices is intuitive. Moreover, even a cursory survey of the history of developmental biology reveals sustained attention to both descriptive and explanatory projects. Consider, for instance, the classic research in developmental biology regarding mapping cell activity and movement. Developing embryos are constituted by cells and cell masses (e.g. cell sheets); moreover, those cells are active (e.g. they migrate, replicate, etc). There is great value in identifying where cells located in some early or earlier stage of development end up in later stages of development. One might, for example, focus on where certain cells of the early gastrula will be found once the basic body plan of the organism is in place. The idea is to construct "fate maps" of cells, and the technique for doing so is to tag cells with long lasting dyes so as to observe their movement. Notice, however, that this sort of project is entirely descriptive, since fate maps work merely to report cell movement. Naturally, developmental biologists have sought also the explanatory basis of events like cell migration; and, as Rosenberg's reductionistic picture contends, it is now commonplace to find a molecular account for things like cell motility.

In this essay, I want to set aside questions about the nature and plausibility of reductionism and focus more narrowly on issues surrounding explanation within developmental biology.[2] In particular, I want to investigate Rosenberg's apparent skepticism about the explanatory potential of pre-molecular developmental biology, as well as his corresponding claim that developmental biology's ability to explain emerges only with its turn to molecular genetics. In contrast to Rosenberg's view, I set out to demonstrate that developmental biology is remarkably consistent (across pre and post the integration of molecular genetics) in both the character of its explanations and its ability to explain. The cornerstone for this position comes

[1] The strong interpretation of Rosenberg's view is supported by selections like the following in his 1997 piece: "At most the non-molecular generalizations set out tasks for developmental explanation, and never provide explanations." (p. 448) To be fair, however, there are occasions where Rosenberg hedges his view, claiming that the embryological level explanations are simply not complete. One might contend that an incomplete explanation is still minimally explanatory, which would run counter to the view that embryological level developmental biology is a wholly descriptive enterprise. The "minimally explanatory" view of developmental biology is much closer to the position I will articulate in this piece, and so for purposes of contrast I'll target the strong interpretation of Rosenberg revealed in the quote above.

[2] Rosenberg deals with a number of traditional concerns surrounding reductionism, including how to understand theoretical terms like "gene" across classical and molecular paradigms. He also makes a point of focus to engage questions of functional terms. I aim to bracket as many of these broad reductionism questions as possible without jeopardizing the issues related narrowly to explanation. For a more direct anti-reductionist critique, though, see Laubichler and Gunter 2001.

with the recognition that explanation in developmental biology is a search for causes, where ‘cause’ is understood as difference-making, and difference-making is established by manipulationist techniques (Woodward 2005). Nevertheless, this revealed uniformity in explanation across developmental biology, both with and without molecular genetics, neglects to account for contemporary developmental biology’s trend toward investigating the molecular genetics responsible for organismic development. In order to reconcile the proposed consistent character of explanation in developmental biology with the preferential focus on molecular genetics, I exploit the notion of *explanatory depth*.

2 Developmental Biology’s Need for Molecular Explanations

Developmental biology, broadly described, is the science that investigates the particulars by which nascent organisms become fully developed organisms; the science might best be summarized by stating that it is the investigation of successive changes of embryos. Like any scientific field, the long history of developmental biology is filled with misguided theory; but once the cell was discovered and the microscope could be put to productive use in the study of embryos, developmental biology centered in large part on the character and behavior of cells that comprised embryos. This embryological level of organization, which is inclusive of the embryo’s individual cells, cell masses or “sheets” of cells, as well as activity of cells and cell masses was the guiding framework for developmental biology prior to the integration of molecular genetics.

For Rosenberg, embryological focused developmental biology proves explanatorily impoverished. Rosenberg’s position regarding the explanatory potential of embryological level developmental biology founds itself on that level’s reliance on theoretical elements that might be described as purely dispositional in nature. Specifically, embryological level developmental biology makes liberal use of elements such as “morphogens”, “organizers” and “inducers”. Ostensibly these elements serve an explanatory role for developmental biology, but Rosenberg works to reveal that that supposed explanatory role is illusory, for those elements lack genuine “empirical content.” Indeed, in reference to developmental biology’s postulating the activity of a “diffusible morphogen” Rosenberg states that this theoretical element, “was advanced as the simplest mechanism to explain certain striking experiments, but there was no independent evidence for the existence of such a substance. At the start, the notion of a ‘diffusible morphogen’ had all the empirical content of Moliere’s ‘Dormative virtue’.” (1997, p. 451) For Rosenberg, neither the “dormitive virtue” nor the analogous diffusible morphogen can be explanatory, presumably because these theoretical elements are specified only by way of their accounting for the explanadum’s occurrence. Put a bit differently, the empirical content of dormitive virtues and diffusible morphogens is exhausted by precisely the state of affairs they are posited to explain.

Rosenberg goes on to insist that theoretical elements like that of the diffusible morphogen merely represent a stand-in for genuine explanations offered at a lower level organization (i.e. molecular genetics). The activity of cells and aggregates of cells must be decomposable into the functional workings of a relatively simply and "decidable" collection of underlying factors, namely those of molecular genetics. In effect, embryological level developmental biology black-boxed developmental explanations by positing theoretical elements identified only by their disposition to produce developmental outcomes needing explanation. Molecular genetics, however, opened the black box(es), and with that developmental biology became an explanatory discipline.

3 The Case for Explanation in Embryological Level Developmental Biology

In reaction to Rosenberg's argument, I want to suggest that the success of his view depends on the adequacy of two additional conditions. The first is that the theoretical elements Rosenberg cites as explanatorily problematic for embryological developmental biology must prove representative of its entire explanatory store. The second is that the characterization of the theoretical elements in embryological developmental biology has to truly run afoul of developmental biology's own presumptive view of explanation. There is, however, good reason to think that neither of these conditions is supported by a proper understanding of embryological level developmental biology. Starting from a causal model of explanation, I want to outline two classic case studies from developmental biology that reveal the explanatory potential of embryological developmental biology. The first case study works to show that the explanatory store of embryological developmental biology extends well beyond that of the elements that worry Rosenberg. The second case study challenges more directly the view that those theoretical elements are non-explanatory.

Before turning to the two case studies, it will be of some value to make a bit more explicit the nature of explanation in developmental biology *as revealed by research practices*. As mentioned previously, I think it can be made fairly clear that developmental biology demonstrates a commitment to explanation through the discovery of causes, where causation is to be interpreted as difference-making, and difference-making is found through the manipulation of developmental conditions. This interpretation of explanation in developmental biology fits neatly within James Woodward (2005) manipulationist framework. The manipulationist conception of causal explanation is manifest in some of the most recognized research findings in developmental biology. Consider, for example, the *Hox-6* gene. *Hox-6* (also known as *antennapedia*) occupies an essential role in *Drosophilia* leg development. This discovery was established quite early in research on *Hox* genes by showing that the gene's inactivity leads to antenna developing in the place of legs, and initiating its activity in the region of antenna development results in legs in place of antenna. Notice how this research strategy conforms to the manipulationist framework insofar as researchers initiated an intervention in the form of inactivating *Hox-6*, and thereby demonstrate *Hox-6* as a candidate part of a causal-explanation for development of *Drosophilia* legs. Experiments like those involving *Hox-6* are standard in developmental biology. But equally important, these explanations are not restricted to developmental biology focused at the level of molecular genetics.

3.1 Case Study 1: The Vertebrate Limb

The vertebrate forelimb is one of the classic points of focus within developmental biology, and it is frequently taken as modeled by the particulars seen in development of the chick wing. It serves additionally as a good first case study for examining the explanatory potential of embryological level developmental biology.

The textbook introduction to vertebrate limb development focuses on three successive developmental events, formation of the stylopod, the zeugopod, and the autopod, respectively. (The stylopod, zeugopod and autopod would in humans correspond roughly to the upper arm, the forearm and the hand.) Furthermore, limb development proceeds from the emergence of a limb bud and the actively of three regions of cells: the apical epidermal ridge (AER); the progress zone (PZ); and the zone of polarizing activity (ZPA). Each of these regulatory regions serves an important role in the development of the vertebrate limb. For the sake of simplicity, though, let us focus on just one: the AER.

The importance of the AER to vertebrate limb development has been demonstrated by the manipulation of chicken embryos. The manipulations, moreover, come in various forms: (1) removing the AER (2) keeping the AER but removing the surrounding ectodermal cell layer on the limb bud and (3) grafting a second AER onto a single limb bud (i.e. there are two separate AERs, the native AER and a grafted AER). In the case of manipulation (1) only a stylopod and the very beginnings of a zeugopod form. In manipulation (2) a fully formed chick wing forms. Finally, in manipulation (3) two fully formed chick wings develop. Even a cursory review of the vertebrate limb reveals an unambiguous illustration of explanation through causal factors. Moreover, the causal model for explaining the vertebrate limb conforms well with a difference-making framework. The explanadum in this case is the developmental outcome of a normal chicken wing (why does the chicken wing develop as it does?). The model can be understood as constituted by various intermittent developmental outcomes (i.e. the stylopod, the zeugopod and the autopod), but most importantly, by the three cell regulatory regions. These cell regions are demonstratively causal elements in the development of the chick wing, as represented by the manipulation of just one of them: the AER. The aforementioned experimental work makes certain the fact that the AER is a component part of the causal explanation of limb development. Moreover, since the AER is representative of how the three cell regions together orchestrate limb development, it is possible to construct a rather robust explanatory model for limb development, a model that restricts its explanatory content to the embryological level of organization.[3]

3.2 Case Study 2: The Chordomesoderm

One of the most astonishing aspects of animal development is not only the diversity of cell types that derive from a pluripotent fertilized egg but also the precise organization of those cells constitutive of a whole organism. One of the foundational processes in this organization is the spatial arrangement of cells that will differentiate into cells that together form an organism's various body parts. Cells destined to become cardiac cells need to be positioned on the interior of the developing organism, while cells destined to be skin cells need be positioned on the outer surface. This organization depends crucially on what is called gastrulation. Gastrulation is the process whereby the developing organism organizes itself with three distinct layers of cells: endoderm, mesoderm, and ectoderm. These three cell layers, in turn, serve to source the cells destined to become the cells of the body parts found in their proper location. In similar fashion, the embryo needs to be organized so as bodily structures form according to their proper location relative to body axes (e.g. anterior/posterior and ventral/dorsal). Now suppose one asks: how do the bodily structures (e.g. the notocord) of an amphibian gastrula get arranged so as to properly develop according to the ventral-dorsal axis? The canonical embryological level answer to this question comes in the form of "the center of organization", or just "the organizer", which is also called the chordamesoderm (Saunders 1982). Of course, for Rosenberg, that answer fails as an explanation because "the organizer", and its associated activity as an inducer, lacks empirical content.

On its surface, the case of the organizer appears to reinforce Rosenberg's concern about the explanatory impoverishment of embryological developmental biology. Nevertheless, closer scrutiny reveals that there is more to the explanatory landscape in examples like the organizer than might first be recognized. Indeed, Rosenberg himself betrays denying wholesale empirical content to his targeted theoretical elements in embryological level developmental biology. In reference to the diffusible morphogen, he states, that it is "a chemical whose concentration gradient would decline as the distance from its source in the embryo, and would switch on different developmental patterns in different parts of the embryo depending on this concentration and the sensitivity of molecular receptors on cell-surfaces or within

[3] It may be worth mentioning at this point that in Woodward's manipulationist framework the idea of a model can come in the form of a "directed graph", which serves as a pictorial representation of proposed causal factors and how the manipulation of those factors affect outcomes. I think explanations in developmental biology could fruitfully be modeled according to a directed graph, but, regrettably, considerations of space make it impossible to produce such models.

them." (1997, p. 451) Rosenberg goes on to state that the problem with the diffusible morphogen is that there "was no independent evidence for the existence of such a substance." But such a claim begs the question regarding the discovery about gradient and associated cell sensitivity to its presence. Fundamentally, there is, I think, a clear tension in maintaining that the diffusible morphogen lacks empirical content and outlining specific empirical findings about the diffusible morphogen (i.e. that it switches on developmental patterns).

The tension seen in the case of the diffusible morphogen is perhaps even more manifest when the question of empirical content is asked of the organizer. There can be little question that developmental biology's recognition of the organizer was an empirical discovery, a discovery famously established by Hans Spemann and Hilde Mangold in 1924. Spemann and Mangold's experimental techniques reveal again developmental biologists' foundational commitment to causal explanation, where the causal elements are revealed by manipulation of developmental systems. Spemann and Mangold's strategy was not only to transplant cells of the dorsal lip between different newt gastrula but to do so between species of newt that had different colored cells. The advantage of that strategy was that it allowed the possibility for tracing descendent cells of those that were transplanted versus those that were host cells. What they discovered is that transplanted dorsal lip cells induce formation of mesodermal bodily structures, principally the notocord, and that most of the cells constituting the notocord were host cells. Spemann and Mangold's assertion was that these dorsal lip cells serve as a central organizer for chordate development insofar as they induce non-neural cells to properly locate and differentiate themselves. Notice that as a consequence of this assertion the inherent display of empirical content surrounding the organizer and its identification as an inducer. The organizer is spatially limited and its activity is temporally (relative to developmental stages) sensitive, for example. Moreover, its discovery as a cause of developmental outcomes demonstrates its explanatory role.

4 The Value of Molecular Explanations in Developmental Biology

The outline of the two case studies above provides a picture of the explanatory character of embryological level. But there are, of course, a variety of ways someone like Rosenberg might respond. One strategy would be to insist on the exceptional nature of explanations in embryological level developmental biology. This argument from exceptions, however, is problematic not only because it effectively concedes the crucial issue regarding a role for explanation in embryological level developmental biology but also because it relies on a dubious presupposition that the two case studies I've outlined are rare. There is a possibility to introduce multiple illustrations of just the sort represented by the vertebrate limb and the organizer. It is tempting, for example, to see embryological level developmental biology's widespread reliance on explanations citing cell adhesion by invoking "cell cement" as equally non-explanatory on the grounds that it is lacking empirical content. But embryological level investigations of gastrulation reveal explanatory power in “cell cement”. Gastrulation involves the invagination of a cell sheet at a particular location of the embryo. This process proceeds by surface cells taking on a shape of bottles, subsequently elongating, and *certain cells more strongly adhering* (via cell cement) to the surface cells that surround the indentation. The result is a pulling of cells surrounding the point of indentation onto the underside of the surface cells, revealing the general procedure for the layering of cell sheets. All of this, again, is spatially and temporally restricted, and so admits of substantive empirical content.

Any argument from exceptions would seem destined to fail, but a second, and perhaps more compelling response to the case studies might look to reestablish directly an inextricable need to cite molecular genetics in explanations of development. At various points in Rosenberg’s argument he makes reference to the notions of explanatory *autonomy* and explanatory *depth*. Unfortunately, Rosenberg never articulates his understanding of either concept; but a successful articulation of these concepts accomplishes, I think, two things. First the distinction better reveals what Rosenberg needs (counterfactually) to restrict developmental biology’s explanations to the molecular level. Second, the distinction supports the case for embryological level developmental biology being an explanatory enterprise while also making sense of the in-general preference contemporary developmental biology has for molecular explanations.

Consider first the notion of explanatory autonomy. Autonomy might be understood loosely here as something to do with self-sufficiency or independence. Rosenberg at various points expresses concern over the completeness of explanations, which might be construed as contiguous with an explanation’s self-sufficiency or independence. Now, in an effort to capture a bit more formally the notion of explanatory autonomy, let me suggest the following. A causal explanation is autonomous just in case the content of the explanans lacks nothing as it concerns their ability to determine the explanadum. This formalization would seem to accommodate the ideas underlying “self-sufficiency”, “independence”, and “completeness”. Moreover, the formalization fits nicely with Rosenberg’s treatment of embryological level “explanations”. Recall, that according to Rosenberg, the problem with theoretical elements like the diffusible morphogen is that it simply serves as a stand-in for content specific, and functionally determinate factors for developmental outcomes. As a consequence, the non-explanatory status of the diffusible morphogen is established, for the diffusible morphogen is explanatorily non-autonomous, since it lacks content that fully determines relevant developmental outcomes.

There is, to be sure, something compelling about the argument that embryological level developmental biology offers non-autonomous explanations, and that molecular genetics provides the resources for fully autonomous explanations of development. I want to suggest, however, that this argument from explanatory autonomy is vulnerable to a regress problem. Perhaps the best way of making clear this regress problem is to return to our first case study, that of the vertebrate limb. In particular, let us now examine a few of the molecular details of vertebrate limb development.

Among a host of other factors, vertebrate limb development relies on four families of protein signaling molecules: fibroblast growth factors (FGFs); wingless-integrative molecular (WNTs); hedgehog gene encoded proteins; and transforming growth factor-βs (TGF-βs). One of the major early findings in the molecular genetics of vertebrate limb development concerned the central role of FGFs. Indeed, FGF-4 appears to be precisely the molecular basis for the signaling between the AER and the underlying cells that proliferate to form the vertebrate limb. This discovery came by way of first removing the AER and then applying FGF-4 to the limb bud; the result of the procedure is proliferation of cells that underlie the AER, leading to outgrowth of the limb (Vogel and Cheryl 1993).

Recognize now the significance of the experiment regarding FGF-4 in vertebrate limb development. First, note that the strategy for determining FGF-4's role conforms to a causal explanation rooted in a manipulationist framework, one that is qualitatively indistinguishable from that of establishing the AER as causally explaining limb development. Second, and more importantly, notice that the charge of lacking explanatory autonomy in the case of the AER applies equally to the explanation provided by FGF-4. The AER is thought to be explanatorily non-autonomous because it lacks content that fully determines why cells proliferate for limb outgrowth. Similarly, though, given only the information presented above, FGF-4 lacks content that determines how it induces cell proliferation. FGF-4 (as is true) may turn out to be a transcription factor. But even if it is known that FGF-4 is a transcription factor, explanatory autonomy would require additional content specifying how FGF-4 promotes or inhibits other gene activity (e.g. the character of the protein constituting the cell receptor for FGF-4 and cascading effects in the cell). Furthermore, the view would seem to require the sequence for the gene that FGF-4 promotes or inhibits, and the functional role that *that* gene plays. As characterized, then, explanatory autonomy would entail a regress, demanding more and more explanatory content. A failure to provide all that content renders an explanation non-autonomous. But the consequence, then, of such a regress would force well-established molecular explanations of developmental outcomes into the category of non-explanatory.

To be sure, some reductionists may look at the argument above and embrace the regress, citing it as support for their view. I have largely set aside the complexities of reductionism to this point, and I have no intent to propose a refutation of such a strong reductionist stance. Instead, I want to simply highlight one of the principal advantages of Woodward's manipulationist framework, namely that concerning its eschewing metaphysics in favor of methodology (2008). As it happens, Woodward does offer an anti-reductionist argument that explanations at lower levels of organization can sometimes be pitched at the wrong "grain".[4] For me, however, I want to emphasize merely how this anti-reductionist argument is subsumed by the general concern regarding making sense of the practice of causal

[4] Woodward outlines the failure in explaining factors related to the temperature and pressure of a gas by appeal to the trajectories of the individual gas molecules.

explanation. Accordingly, developmental biologists *in practice* seem to successfully explain quite frequently, and, historically, they have done so at different levels of organization. In this light, the regress issue focuses our attention on two inter-related matters. First, there is the question of whether an appeal to practice is a successful strategy for defending the autonomy of embryological level explanations. Second, there is a need to examine Rosenberg's exploiting the fact that explanatory models often leave explanatory gaps.

Consider first the issue surrounding developmental biology's practice of explaining developmental events and outcomes. Rosenberg acknowledges the appeal to practice as an available anti-reductionist strategy, but he denies its success. In his earlier piece (1997) he compares classical genetics to Newtonian mechanics insofar as both are retained for heuristic value. In the later work Rosenberg responds to a position he dubs explanatory Protagoreanism—"the thesis that 'some human or other is the measure of all putative explanations, of those which do explain and those which do not.'" (2006, p. 35) As our two case studies demonstrate, however, neither of these arguments is representative of the practice of explanation in developmental biology. Embryological level explanations of development are not mere limiting cases of molecular explanations as is seen in Newtonian mechanics (i.e. they are not true in some special circumstances but not in the whole range of relevant conditions). Nor is the appeal to practice a license to admit anything as a successful explanatory model. Indeed, the "anything goes" view of explanation would allow a kind of non-responsiveness to empirical findings. But it's clear that the manipulationist framework puts empirical results at the very center of explanation.

The above suggests that the foundational issue in understanding embryological versus molecular explanation in developmental biology lies not with the notion of explanatory autonomy but something else entirely. An autonomous explanation, within a manipulationist framework, is one that conforms to the demands of successfully revealing causal relationships that allow model construction; moreover, models must be responsive to empirical investigations involving manipulation of proposed causal factors. Contrary to Rosenberg's position, the preferential shift to molecular explanations in developmental biology was not a function of embryological level developmental biology proving a non-explanatory, purely descriptive enterprise, for the case studies outlined previously demonstrate how embryological level developmental biology built models that satisfied this demand. The real issue, therefore, appears to turn on how to treat the aforementioned explanatory gaps left by autonomous explanatory models.

One of our case studies is again instructive. The organizer proved a revolutionary discovery in developmental biology, but by the 1970s (some 50 years later), the organizer was of little interest to working developmental biologists. In sum, derivative explanations stemming from the organizer were not forthcoming. Developmental biologists sought greater understanding of developmental organization, but just as Rosenberg rightly shows, that understanding demanded articulation of the relevant molecular genetics operating in cell organization and induction. Recognize, then, the shift to molecular explanations in developmental biology was not a shift from an inability to explain to a capacity to explain; rather it signaled developmental biology's need for increasing its explanatory depth.

The notion of explanatory depth has been given some attention within another difference-making analysis of explanation, that of Michael Strevens (2009). Strevens' account of causal explanation differs markedly from that of Woodward, particularly with respect to the degree to which explanation requires moving to lower levels of organization. Strevens is much more demanding on this front, advancing the view that it is "compulsory" that explanatory models seek greater depth.[5] Like Rosenberg, Strevens cites the example of the dormitive virtue as illustrating the need for explanatory depth. Predominantly, though not exclusively, Strevens views depth as determined by the level of organization at which one explains. But perhaps more important in this context is the significance of Strevens' linking the practice of causal explanation to that of understanding. There is significant plausibility, not to mention value, in seeing an increase of understanding (via explanation) as a function of the complexity—measured by the number of proposed causal factors—of an explanatory model. The reason, as we've seen with the comparison between embryological and molecular explanations is that higher levels of organization will typically leave causal or explanatory gaps between intermediate states and events specified in a higher level model. As explanatory depth increases, though, and one constructs models at lower levels of organization, one expects to identify a larger set of causal elements, which, in turn, causally relate intermediate states and events otherwise causally unconnected. The consequence that emerges from this sketch of explanatory depth is a view that reconciles the explanatory potential of embryological level developmental biology with the undeniable trend in contemporary developmental biology to investigate the role of molecular genetics in development. In the end, explanations within embryological developmental biology are autonomous, but they are typically not as deep.

References

Laubichler, M., & Gunter, W. (2001). How molecular is molecular developmental biology? A reply to Alex Rosenberg's reductionism redux: Computing the embryo. *Biology and Philosophy, 16*, 53–68.

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: University of Chicago Press.

Saunders, J. W. (1982). *Developmental biology: Patterns problems principles*. New York: Macmillan.

[5] Strevens states, however, that this requirement does not entail explanation is forced down to the level of elementary particles. Like Woodward, he cites the example of the conditions of a gas as illustration of stopping at a higher level of organization.

Strevens, M. (2009). *Depth: An account of scientific explanation*. Cambridge: Harvard University Press.

Vogel, A., & Cheryl, T. (1993). FGF-4 maintains polarizing activity of posterior limb bud cells in vivo and in vitro. *Development, 119*(1), 199–206.

Woodward, J. (2005). *Making things happen: A causal theory of explanation*. New York: Oxford University Press.

Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research, 77*(1), 193–212.

Synthetic Genomics and the Causal Role of Genes: What has been Shown and Why it Matters

Bettina Schmietow and Lorenzo Del Savio

Abstract Synthetic genomics, the synthesis of a whole viable genome of a self-replicating organism, has been mainly considered as a momentous technical achievement. Yet the creation of a bacterial cell from a synthetic genome has also a considerable scientific significance, insofar it tackles a crucial question of the genomic paradigm of research: "Do chromosomes contain the whole genetic material?" In this paper, we situate the synthetic cell experiment in this tradition of research and explore the conceptual consequences of the resulting empirical findings. We argue that a technological understanding of synthetic genomics is partial and that it must be supplemented with a discussion of the notion of *artificiality*, that depends on the claim that "the DNA software builds its own hardware", and which is the case if and only if the DNA contains the entire genetic repertoire of an organism.

1 Beyond a Technology-Only Understanding of Synthetic Genomics

In July 2010, the journal *Science* published the first research article that described the successful creation of a bacterial cell controlled by a genome that had been chemically synthesized (Gibson et al. 2010, pp. 52–56). An artificial genome polymerized after a genomic sequence of the simple prokaryotes *Mycoplasma*

B. Schmietow (✉) • L. Del Savio
European Institute of Oncology, University of Milan, IFOM-IEO campus,
via Adamello 16, 20139 Milan, Italy
e-mail: bettina.schmietow@ieo.eu; lorenzo.delsavio@ieo.eu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_29,

mycoides[1] was implanted in a recipient cell-like vesicle provided by a *Mycoplasma capricolum* donor, a bacterium belonging to the same genus. Resulting cells were afterwards proven to be viable and phenotypically indistinguishable from natural *M. mycoides*. The experiment was the first successful integration of two technologies that had been developed at the Craig Venter Institute (JCVI): whole-genome synthesis and genome graft.

Following extensive media coverage of the experiment and the "preventive ethics" strategy at the Craig Venter Institute (Cho and Relman 2010, pp. 38–39), societal consequences of synthetic genomics have been widely explored. The main approach to the latter is epitomized by the comment made right after the announcement of the experiment by Mark Bedau in *Nature*.

> The "synthetic cell" created by Craig Venter and his colleagues is a normal bacterium with a prosthetic genome. As the genome is only about 1 % of the dry weight of the cell, only a small part of the cell is synthetic. But the genome is pivotal because it contains the hereditary information that controls so much of a cell's structure and function. (Bedau 2010)

According to Bedau, Venter's achievement is mainly a technological one: before the experiment, we had only been able to modify genomes in a piece-meal fashion through ordinary genetic engineering; after that experiment we can take a whole-genome approach, having "*a proof of principle for producing cells based on computer-designed genome sequences*" (Gibson et al. 2010, pp. 55). Ethical issues have thus been thus tackled along this *merely technological* understanding of synthetic genomics. Unsurprisingly, the conceptual continuity of these techniques with traditional genetic engineering was supported in several papers (Cho and Relman 2010, pp. 38–39; Cho et al. 1999), thus leading to the general acceptance of the technology as a moral *fait accompli* – in fact resolved several years before synthetic genomics (Balmer and Martin 2008; Boldt and Mueller 2008).

We argue that the technological understanding of synthetic genomics is partial and that it must be supplemented with a discussion of the notion of *artificiality*, so far easily dismissed with remarks on the naturality of the cytoplasm of the donor cells or the use of the term "prosthesis", which entails that only parts of the organisms were artificial. These remarks fail to appreciate that creating a wholly artificial cell was the *scientific* aim of the experiment. In the interpretation of the scientists at the JCVI, that aim was indeed successfully fulfilled:

> The properties of the cells controlled by the assembled genome are expected to be the same as if the whole cell had been produced synthetically (the DNA software builds its own hardware). (Gibson et al. 2010, p. 56)

[1] The sequence was substantially similar to that of naturally occurring *M. mycoides*. It also contained four watermark sequences that were introduced to tell apart the synthetic organisms from their natural counterparts.

Moreover, that artificiality ought to be at the center follows from the crucial question addressed by the paper:

> Even in simple bacterial cells, do the chromosomes contain the entire genetic repertoire? If so, can a complete genetic system be reproduced by chemical synthesis starting with only the digitized DNA sequence contained in a computer? (Gibson et al. 2010, p. 52)

The claim that the DNA software builds its own hardware has some support only if the authors are able to argue that DNA contains the entire genetic repertoire. If the scientific claims of Gibson and colleagues are to be taken seriously, we ought to focus on the issue of artificiality since this would remain the radical novelty of their research program aside from the refinement of pre-existing technologies for the modification of genomes.

So far, whatever living organism we have encountered belonged to one of the branches of our common tree of life. Some of these organisms are more or less heavily modified by humankind, e.g. the ones that are bred to serve our purposes and the ones that are interfered with by means of bioengineering. If Venter and colleagues' claims are sound, however, we will be likely to live along with new kinds of organisms that stem from computers and in fact, the *intelligent design* of their creators. This is why the artificiality-centered understanding of synthetic genomics leads to a richer discussion of the societal consequences of that enterprise.

In this paper, we develop the details of an artificiality-centered understanding of synthetic genomics by pursuing a conceptual clarification of the notions "synthetic cells" and "genomes". To this end, we will draw on some philosophical resources that had been refined within the debate on the causal primacy of genes.

The plan of the work is as follows. In Sect. 2, we take advantage of a distinction made by Barnes and Dupré (2008, pp. 75–109) between genomes as information and genomes as matter in order to tell apart two threads of synthetic genomics research. Along with scientific papers, further resources will be patents on synthetic genomics, which are particularly relevant to understand the overall strategy of the research program. We will argue that the bacterial cell synthesis was set up to join together two parallel branches of research at the Craig Venter Institute, biochemistry of DNA synthesis and genetics. In Sect. 3, we briefly describe the basics of Gibson and colleagues' experiment in order to lay the foundations for the subsequent discussion. In Sect. 4, we challenge an interpretation of the results of the experiment that was defended by Daniel Dennett (2011) and that converges with the one by Gibson and colleagues themselves. In particular, we analyze some objections that can be posed against their thesis: the inadequacy of the use of a recipient cell which is philogenetically close to the donor genome to make a point about the exclusive role of genomes in determining the phenotype of the organism; and the critique that is based on difficulties that scientists encountered regarding the peculiar causal role of some non-genetic mechanisms that can be understood to have an informational role as well. To conclude, in Sect. 4, we will sum up the main conceptual results of the paper, calling for an in-depth consideration of these scientific aspects of the methodology in the public debate about these experiments.

2 Genome as Information and Genome as Matter

Barnes and Dupré (2008, pp. 75–109) distinguish three different conceptions of genomes that appear in the literature, i.e. in scientific handbooks and papers.

(a) *Genome as matter*: there are two material genomes in each somatic cell of our body. Genomes as matter are *made of DNA*: they are those nucleic acids that, upon mitosis, condense in the 23 chromosomes within the nucleus.
(b) *Genome as a set of genes*: there is just one genome as set of genes for each person. Genomes as set of genes descend from an understanding of genetics that is by now surpassed: genes are discrete stretches of DNA that are transcribed and translated into proteins.
(c) *Genome as information*: this is a refinement of conception (b) and even in this case there is just one genome for each person. That is, each person has her own genomic sequence, which is implemented on molecules of DNA in each cell of our bodies.

There are interesting relationships between these conceptions: (b) is just a subcase of (c) and in fact it refines the latter in order to take into account some complications that challenge the traditional definition of genes. Conceptions (a) and (b) *denote* different entities whose relationship it is worthwhile to explore empirically. Using these conceptual tools, we can express the aim of Venter's experiment as such an empirical exploration. To support this idea, we analyze two patents that were filed by scientists working at the institute, one involving genomes as information and the other genomes as matter. These documents provide a thorough description of two different threads of research carried out by synthetic geneticists that were conjoined in Gibson et al.'s experiments.

As for the genome as information, it is informative to consider the patent on the "Minimal Bacterial Genome" (WO 2007/047148 A1). The search for a minimal organism characterizes several branches of artificial biology since its inception. There are bottom-up approaches that try to boot up life from non-living manner, for which a "minimal organism" is just a chemical system that has relevant properties and hence deserves to be called "life". On the other hand, there are top-down approaches that start with complex entities and try to eliminate as much as possible in order to single out the minimally complex organism.

The notion of complexity, however, is famously intractable. The pragmatic approach taken at the Craig Venter Institute was to start with organisms that are considerably simple according to the same measures and then to make them even simpler by interfering with their structure. *Mycoplasma* has been the class of organisms of choice because, even though it is undisputable that *Mycoplasma* are forms of life, they were selected throughout their evolution for small dimensions and short genomes as a consequence of their parasitic life. As the authors observe in the patent, "*Mycoplasma genitalium is already close to being a minimal bacterial cell*": it contains only 482 protein-coding genes (compare it with ca. 20.000 coding genes in humans), where a reasonable number of essential genes obtained by comparison of phylogenetically distant bacteria sums up to 256.

The simplification procedure that was carried out on *M. genitalium* was strongly biased toward the understanding (b) of genomes: genomes are set of genes. Even though that conception lacks precision in general, it still retains theoretical relevance in simple prokaryotic life, where the conception was firstly developed and then transferred to other forms of life. As François Jacob once said, what is true for *E.coli* is also true for an elephant (Jacob 1993). Starting from the conception (b), simplification could mean only reduction in number, that is, a smaller set of genes is simpler. Hence, the researchers targeted genes by mutagenesis, one by one, to select those that seem to be non-essential for viability. They were able to eliminate 101 genes and to retain 381 in the set of the minimal bacterial genome (in a specified stress-free environment).

Minimal bacteria would then be used as a scaffold to attach functions that are relevant for human purposes. One missing step, however, is the production of organisms that comprise scaffold and functional parts. In this respect, a second patent is worth-analyzing: Patent WO 2008/24129 A2, which covers methods for constructing synthetic genomes, comprising designing DNA sequences, generating synthetic DNA molecules and transplanting them into cells.

The patent describes large parts of the protocol of the synthetic bacterial cell experiment. Here, the term "genome" denotes a piece of matter, the one that is synthesized out of four nucleotides. The patent contains also a specification of the term "synthetic genome", which is considerably different from the one provided above:

> A "Minimal replicating synthetic genome" is a single polynucleotide or group of polynucleotides that is at least partially synthetic and that contains the minimal set of genetic sequences for a cell or organelle to survive and replicate under specific environmental conditions. (Patent WO 2008/24129 A2)

This definition is of interest for our analysis that goes beyond the distinction between informational and material genomes. It indeed hints at the problem that will be tackled in the experiment by Gibson and colleagues. Does the material genome contain all the information that specifies an organism that is able to "survive and replicate under specific environmental conditions"? is the *informational* genome just a string of A, C, T and G that stand for nucleotides or should it comprise additional information?

Their main hypothesis, which also lies behind the idea that a prosthetic genome is enough to render the whole cell artificial, is the following: the informational genome, at least in simple bacteria, is indeed a string of A, C, T and G. Furthermore, its material implementation, when loaded onto the correct receptacle, is able to specify the replicating system as a whole. As they formulate quite imaginatively in the paper, *the DNA software builds its own hardware* (Gibson et al. 2010, p. 56).

In order to understand what this hypothesis entails it is worth examining two passages of the patent, which describes the role of the environment and the non-genomic parts of the cell in a replicating system:

> Of course, nutritive, metabolic and other substances as well as physical conditions such as light and heat may be provided externally to facilitate the growth, replication and expression of a synthetic cell. (Patent WO 2008/24129 A2)

> The cellular genome is supplemented in the vesicle (e.g., cell) [...] with complex components such as ribosomes, functional cell membranes, etc. These additional elements may complement or facilitate the ability of the genome to achieve (e.g. program) replication of the vesicle/cell. (Patent WO 2008/24129 A2)

According to this hypothesis, although external environment and parts of the cell (i.e. internal environment) are meant to facilitate replication, they do not have any *informational* role. Notwithstanding their necessity for growth and replication, they do not make a difference as for the resulting phenotype of the organism. Difference-making in phenotypes seems indeed the informal understanding of the authors concerning what it takes to have an informational role in development and inheritance, this notion being couched in terms of opposition to the "merely supportive" role of the environment. Moreover, it should be noticed that the informational metaphor can be dropped altogether, and the same question put more neutrally in terms of the relative causal contribution of genes and other parts of the cell as for the process of development and inheritance.

Their hypothesis consists in positing that external and internal environment are simply a receptacle on which all kinds of genomic instructions could be booted up. These instructions specify the phenotype of the organism, which is therefore wholly contained in the chromosome as in a string of A, C, T and G within the memory of a computer. In philosophical debates, the latter thesis has been sold as "gene determinism", which philosophers mostly assumed to have a single answer. Here, the same thesis has been made investigable by narrowing down its scope to the small class of organisms *Mycoplasma*.

As a consequence, the experiment described in Gibson et al. (2010) is a test of gene determinism as far as Mycoplasma is concerned. In the following, we will describe the test, an interpretation favored by Dennett and the scientists who carried out the work, before adding a couple of skeptical remarks.

3 Instructions for the Creation of a Bacterial Cell

How would you test whether chromosomes contain the whole genetic repertoire? Equivalently, how would you test whether a string of A, C, G and T does specify a replicating system or whether internal and external environment are a sheer receptacle? There have been many experiments, in several types of organisms, which took up this very same question. A clear answer to the problem should be given indeed by whatever experiment that has the following form: take two organisms, switch their genome, and check whether they switch also their phenotype. Seminal experiments by Ayers in the 1944 paper on *Pneumococcus* relied on this experimental pattern, and that is the case even with the tobacco mosaic virus experiments in the 1950s (Morange 1998, p. 62). Cloning species using donor oocytes from different species also implement this form of experiment. Venter's lab followed the same approach with the relevant difference that the genome they used was *artificial*.

The idea developed at the CVI is roughly the following: if, by taking two species of cells, *M.capricolum* and *M.mycoides*, we are able to transfer the genome of the latter in the cell machinery of the former and the result is a new system that is phenotypically indistinguishable from *M.mycoides,* then we have indeed shown that the cell machinery, just like the environment, is functioning merely as a receptacle. Furthermore, if we also synthesize the genome starting with a sequence contained in a computer, then we really have a proof of principle that it is the bare sequence that performs the specification and not any further features of the material genome.

Simple as it may seem, the realization of this experiment took several years. Two technical obstacles had to be overcome. First, the transplantation of an alien genome into a cell was and remains very inefficient. Second, the synthesis of the long molecule of genomic DNA must be very precise, up to single nucleotides. As for the latter issue, it is worth recalling that the experiment was stopped for several months because of a point mutation in a single gene essential for viability, which gives an idea of the precision of synthesis required for a successful experiment.

The approach to the synthesis was stepwise: short stretches of DNA were glued together, cloned and moved up to next past-and-clone step. The shortest cassettes were 1,000 base pairs long (plus 80 bp overlapping with the continuous cassettes to paste them together) and produced with chemical methods. This was about the longest stretch of DNA possible to generate chemically with a low probability of mistakes. The final genome is 1.077.947 bp long and the assemblage required three steps (1-10-100-1.000 kbp). Each step involves the recombination of the cassettes, their cloning and selection. The first step has been made in *E.coli* while the second and the last are performed in yeast, the sequence to be cloned being too long to be managed by the *E.coli* copying machinery.

The genome was then sequenced and proofread. Upon transplantation and selection (synthetic genomes were tetracycline resistant), synthetic cells were obtained and scored for similarity with wild type *M.mycoides* by proteomics analysis. The phenotype of the synthetic cell does resemble the WT *M.mycoides* and not *M.capricolum*. The sequence that was used had been obtained from a *M.mycoide*s reference sequence that underwent some modifications. Of particular interest as for the protocol were the "watermark" sequences used to distinguish artificial genomes: they contain specific restriction sites that, upon enzymatic digestion, allow an easy recognition of *M.mycoides* JCV1.0. Furthermore, sites for selection were added: a tetracycline resistance cassette and a protein-coding sequence which produces a blue pigment upon exposition to X-gal. The genome is therefore synthetic in a meaningful sense only if we assume the material understanding of genomes, and indeed the sequence is almost entirely identical to naturally occurring bacteria, aside from a couple of protein-coding genes necessary for the selection.

Of great interest for our discussion to follow is a modification that had to be made to recipient cells. Bacteria contain restriction enzymes that cut DNA in a site-specific manner. This machinery is supposed to be a protection system against viral invasion. The sequences that are recognized by the enzymatic system are, however, quite widespread even in the resident genome. Protective mechanisms had to evolve in order to prevent a sort of rudimentary autoimmune effect with lethal consequence.

The protection usually consists in the heavy methylation of the sequence recognized by the enzyme, that is, the addition of chemical moieties (a methyl group) to cytosine in the DNA.

4 Discussion: Does the Life Software Build its Own Hardware?

Intervening in a debate on whether information plays (or should play) a role at all in the life sciences, Daniel Dennett proposes a thought experiment. Two persons decide to have a child in a tortuous manner: they have their genomes sequenced, apply onto the sequences thus obtained a meiosis algorithm and then synthesize the newly designed genome in order to implant it. Dennett claims that there is a *certain* sense in which this child is the biological offspring of their parents (Dennett 2011). Nobody would probably deny this fact. His bemusing conclusion is that, since the genome was passed through an information-processing device, we have a proof that inheritance is a matter of transmission of information. We will not comment here on the whole conclusion, but would like to discuss whether it is the case that if we are able to obtain a new organism through a computer bottle neck, then what matters for the transmission of information is only the genomic sequence. As Dennett puts it:

> It is now possible to take the information and use it to construct a new vehicle for that information that can be read just fine by the organism that contains it. (Dennett 2011)

How do we know that none of the features of the newly born baby are due to the donor cell in which, eventually, the genome was implanted? In this respect, Craig Venter's experiment is much more refined than Dennett's thought experiment. In fact, they used a donor cell from a *different species* and checked whether the resulting phenotype was more similar to the donor of the cytoplasm or to the donor of the genome, concluding in favor of the latter.

There does, however, remain a conceptual issue to be addressed: why did Gibson and colleagues decide to use a donor cell of *M.capricolum* rather than one from a more distantly related bacterium? If their question was whether the genetic repertoire is wholly contained in the chromosome, it seems that it would have been conceptually sharper to go for a more distant organism as for the donor of the "reading-machinery". In fact, if, upon protein turn-over, even the phenotypic effects of these radically different cells had faded away, we would have a clearer proof that only genomes matter as for the specification of the organism.

This proposal, of course, is biologically naive, for such an experiment would probably result in unviable cells. One could argue that this would be as if cells were deprived of the right kind of nutrients. Nobody would go for the impervious road that the genetic repertoire is contained even in the nutrients or in the physical environment on the cell. For the same token, we need the right kind of reading machinery for cell viability. That is, one might protect the argument by recurring to a sophisticated distinction between viability and *forms* of the cell, the first granted by the right kind of facilitating internal and external environment and the latter specified by the informational molecule.

In addition, the experiments give some reason to think that, even if we were able to provide the distinction above, we would anyway not be able to replicate the results using different organisms. The donor cell was not a wild type *M.capricolum*, but rather a *M.capricolum* that lacked the restriction enzymes that would have cut the unmethylated artificial genome. This is only the simplest of the hurdles that epigenetic modifications of the genome create to synthetic genomics. In higher organisms, different kinds of epigenetic modifications are known and while some may be only hurdles for viability, others are established to be crucial for the specification of certain phenotypes. It seems, in other words, that while rendering the question empirical, they also made the answer so narrow to be of no theoretical interest as for the role of genomes in specifying organisms in general.

This latter question lies at the core of the traditional debate on the causal primacy of genes. The received view on the issue – gene-centrism – is the conjunction of two theses: (1) genes are the only units of *inheritance* and (2) genes *determine* the development of an organism (in the case of multicellular organisms) or its structure (in the case of single-celled organisms). These propositions are intertwined. *Inheritance* (1) is simply the fact that the like begets the like, a fact that is commonly explained by a theory of transfer of causally crucial material from parents to offspring (Mameli 2005), where this causal role is spelled out in terms of *determination* (2). Gene-centrism has mainly been challenged in two different ways. Either it is argued that the notion of a gene is blurred, or other factors in addition to genes are shown to be transmittable from parents to offspring (Jablonka and Lamb 2005), causally determinant in development (Oyama 1985), or both (Moss 2003). The second family of objections is particularly interesting because it suggests empirical inquiries, namely whether there are other inherited materials beyond deoxyribonucleic acids that determine the organization of an organism.

Notably, the paper by Craig Venter's team reports an experiment that was explicitly set up to address empirically the gene-centrism debate in the case of *simple bacterial cells* and heavily relies on informational talk for the discussion of the results. Venter's main tenet is simple: if the DNA sequence (genome) of an organism A contains its genetic *repertoire*, then we should be able to sequence it, store it in a computer, chemically create a synthetic chromosome with the same sequence and obtain an organism A by implanting the synthetic chromosome into a chromosome-depleted organism B. Substitute A with *Mycoplasma mycoides* and B with *M. capricolum* and you will get to Venter's teams' experiment. The result has been described above: if a donor synthetic chromosome that was derived from a sequence of *M. mycoides* is implanted in a recipient *M. capricolum* cell, the latter reverts its phenotype with protein turnover and becomes eventually indistinguishable from the former. According to the authors, this proves that (1) the DNA sequence was accurate enough to specify a viable organism; (2) "DNA sequencing of a cellular genome allows storage of the genetic instructions for life as a digital file"; and (3) "the DNA software builds its own hardware". In other words, we have empirical evidence that gene-centrism is true in the case of *Mycoplasma*.

As we have tried to show, this thesis is disputable. To begin with, other factors are needed to obtain viable cells, namely whole donor cells. One cannot boot up a cell from scratch: though DNA contains the genetic repertoire, a naked DNA cannot pull itself up from its own sequence. A straightforward reply consists in pointing out that those further factors are not specific: without them, cells are simply not viable. It is not the case that using factors derived from other organisms one obtains cell types specific of that organisms. A further and more technical objection relates to the fact that Venter's team actually had to interfere with the donor cells in order to create viable *M. mycoides*. Donor cells were depleted of some restriction enzymes that cut foreign DNA in a sequence-dependent manner: endogenous DNA is normally heavily methylated around these restriction sequences whereas synthetic genomes are not and hence would be digested. Though DNA-methylation is known to carry indispensable developmental information in eukaryotes, defenders of gene-centrism might reply that in the case of bacterial cells this objection boils down to the one above, already rejected through the specificity vs. viability counter-argument. Finally, the experiment was shown to work using very closely related organisms: it might be argued that one would not obtain *M. mycoides* cells using less closely related species or that, in any case, this is an open empirical question that requires an answer before we can claim that we have a proof of principle for storing life as a digital file that contains a DNA sequence. On the information side, Dennett claimed that the fact that an information-processing device was an intermediate in the passing on of genetic material during the creation of synthetic cells proves that inheritance is "fundamentally an information-transmission process" (Dennett 2011). Yet, it is not clear that this unusual route of transmission may lend support to any side of the debate.

To wrap up, we have argued that Venter's experiment sheds new light on the debate on gene-centrism. Moreover, we have shown that this conceptual issue provides the foundation for the discussion on the artificial nature of the bacterial cells created by synthetic genomics, that is, whether a cell whose synthetic core weights about 1/100 of the total dry-weight could be considered man-made. This issue should be explored further in order to evaluate the continuity or discontinuity of this new set of techniques with traditional genetic engineering.

References

Balmer, A., & Martin, P. (2008). *Synthetic biology: Social and ethical challenges*. Nottingham: Institute for Science and Society, University of Nottingham.

Barnes, B., & Dupré, J. (2008). *Genomes and what to make of them*. Chicago: The University of Chicago Press.

Bedau, M. (2010). Life after the synthetic cell. *Nature, 465*, 422–424.

Boldt, J., & Mueller, O. (2008). Newtons of the leaves of grass. *Nature Biotechnology, 26*, 387–389.

Cho, M. K., & Relman, D. A. (2010). Synthetic "life", ethics, national security, and public discourse. *Science, 329*, 38–39.

Cho, M. K., et al. (1999). Ethical considerations in synthesizing a minimal genome. *Science, 286*, 2087–2090.

Dennett, D. C. (2011). Homunculi rule: Reflections on Darwinian populations and natural selection by Peter Godfrey Smith. *Biology and Philosophy, 26*, 475–488.

Gibson, D., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science, 329*, 52–56.

Jacob, F. (1993). *The logic of life: A history of heredity*. Princeton: Princeton University Press.

Jablonka, E., & Lamb, M. J. (2005). *Evolution in four dimensions. Genetic, epigenetic, behavioral, and symbolic variation in the history of life*. Cambridge, MA: MIT Press.

Mameli, M. (2005). The inheritance of features. *Biology and Philosophy, 20*, 365–399.

Morange, M. (1998). *A History of Molecular Biology*. Cambridge, MA: Harvard University Press.

Moss, L. (2003). *What genes can't do*. Cambridge, MA: MIT Press.

Oyama, S. (1985). *The ontogeny of information: Developmental systems and evolution*. Cambridge: Cambridge University Press.

Part X
Philosophy of the Life Sciences: Biological Knowledge and Structural Realism

Eschewing Entities: Outlining a Biology Based Form of Structural Realism

Steven French

Abstract Structural realism finds its natural home in physics. Nevertheless, I argue that a form of structural realism can be elaborated in the context of the biological sciences as well. In particular, just as there is a problem with individuality in quantum physics, so one can argue that there is a problem of biological individuality that motivates a move away from biological entities as elements of our fundamental ontology in this area. Here I sketch some of the implications for this view and respond, briefly, to the claim that the use of genes as 'levers' in biological practice supports an object-oriented stance towards them.

1 Introduction

Both the elaboration and criticism of structural realism have typically been articulated in the context of physics. Structural realism is motivated by, first of all, the presence of mathematical equations that allow straightforward representation of the relevant structures; and secondly, the implications for the individuality and identity of putative objects. My aim here is to explore the possibility of developing similar views in the biological domain. An obvious concern is that within these contexts we may not be able to find highly mathematised structures. Elsewhere I have indicated how the representational framework of the model-theoretic approach might help allay such concerns (French 2011). Furthermore, issues of object identity and individuality arise here as well. Thus, Dupré insists that there exists a 'General Problem of Biological Individuality' which concerns the issue of how one divides 'massively integrated and interconnected' systems into discrete components. His

S. French (✉)
School of Philosophy, Religion and History of Science, University of Leeds, Leeds, LS2 9JT, UK
e-mail: s.r.d.french@leeds.ac.uk

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_30,

solution is to advocate a form of ‘Promiscuous Realism’ that holds, for example, that there is no unique way of dividing the phylogenetic tree into kinds. Instead I have urged serious consideration of those aspects of the work of Dupré and others that lean towards a structuralist interpretation (ibid 2012). Here I want to suggest further possible ways in which a structuralist stance might be developed within biology.

2 Laws and Symmetries

Let us begin by recalling the twin motivations for structural realism in responding to theory change and to the ontological implications of our theories with regard to putative objects. With regard to the former, the response of the structural realist is to uncover the structural ‘commonalities’ between the relevant theories and urge the realist to place her ontological emphasis on these. Using examples from physics, such commonalities are typically identified via the relevant equations and/or laws purportedly carried over from one theory to its successor – as in the classic case of Fresnel’s equations, recoverable from Maxwell’s, for example. With regard to the latter, the metaphysical invariance typically associated with objects is shifted onto the relevant symmetries that, in physics, yield both the kinds of particles (as permutation symmetry does for bosons and fermions) and their properties (thus the space-time symmetry captured via the Poincaré group yields mass and spin, for example). Indeed, in this context the notion of ‘structure’ can be cashed out in terms of these laws and symmetries which at the level of theories themselves are presented via the relevant mathematical formulation (group-theoretic in the case of the symmetries) and at the meta-level of the philosophy of science by the set-theoretic structures of the semantic approach, for example. These laws and symmetries can be ‘read off’ our theories in realist fashion and the relevant properties are then identified in terms of the role they play in these laws and symmetries. At that point the structuralist urges that we stop and do not make the further move of taking these properties to be possessed by objects. It is the laws (and symmetries) that we take as representing the structure of the world (French forthcoming).

But if that is what is meant by ‘structure’, then there would appear to be obvious obstacles to articulating a similar stance within the biological context.

Thus, although it might seem that the kind of broad correspondence underlying the above claim of commonality could also be claimed to exist in the biological domain – think, for example of the claim that chromosome inheritance theory reproduces Mendel’s laws of inheritance (where it is granted that the inherited factors are not quite as Mendel conceived them; this being analogous to changes in our understanding of the underlying nature of light, say) – in biology we face the obvious problem of a comparative paucity of mathematised equations or laws by means of which we can identify and access the relevant structures.

But of course there are lots of structures in biology, presented via the models of the relevant theories and which can be represented at the meta-level by the semantic approach (see Odenbaugh 2008). Nevertheless, there remains a major contrast with physics, for example, in that these models describe the contingent outcomes of evolution (Beatty 1995). Now one option is just to bite the bullet and accept that these models and the associated biological generalizations are fundamentally evolutionary, in the sense that under the effects of natural selection they themselves will evolve.

In this sense, they cannot be said to hold in all possible worlds and thus cannot be deemed 'necessary', in the manner in which standard examples of laws taken from physics are regarded as necessary, as opposed to accidental. If lawhood is tied to necessity, then such generalizations cannot be regarded as laws. However, given their role in biological theory, they cannot be dismissed as mere accidents like the claim that I have 67pence in my pocket. They have more modal resilience than that (cf. Mitchell 2003). Putting this resilience together with their evolutionary contingency in the structuralist framework yields a form of 'contingent structuralism' in the sense that, unlike the case of physical structures where the structural realist typically maintains that scientific progress will lead us to *the* ultimate and fundamental structure of the world, biological structures would be temporally specific, changing in their fundamental nature under the impact of evolution.

Now a striking feature of the invariant biological regularities represented by models is the variety and heterogeneity of the limitations imposed upon them. Rosenberg puts the point nicely as follows:

> … once environments come to include creatures and their effects on one another, the lifetimes of regularities about creatures' adapted traits fall from the scale of billions of years (Archea—whose environment has not changed for 3 billion years) to multiple geological epochs (oxygen-respirators) to hundreds of millions of years (vertebrates) to weeks and months in the case of others (the AIDS-virus). (Rosenberg 2011, pp. 11–12)

Consider the following example (which exemplifies the typical form of laws considered by philosophers): 'All genes are composed of DNA'. As Rosenberg notes, over a long period, this regularity remained invariant but by virtue of being subject to no exceptions, '… its operation provided an environment that would allow for the selection for any new biological system that could take advantage of the fact that all genes are composed of DNA.' (ibid, p. 12 fn. 11) Of course, such a system eventually evolved, namely RNA viruses, which parasitize the machinery of DNA replication (as an example, consider the HIV virus). Thus what Rosenberg calls the arms race of evolutionary competition generated a shift from 'All genes are made of DNA' to 'All genes are made of nucleic acids (either RNA or DNA)', with further shifts possible in the future.

Now in order to explain this variety of limitations on invariances, Rosenberg argues, we need to appeal to laws and in the biological domain the laws required are those of natural selection. It is in this manner that we can explain the differences in both the limits and success of models. So, for example, in the case of the Lotka-Volterra model the invariance is broader than that exhibited by Nicholson-Bailey models of bacterial parasites and hosts. More importantly, there are no spatio-temporally unrestricted regularities in biology, something that depends on the laws of natural selection. Thus, the principles of Darwinian evolution would have to be regarded as nomological generalizations of the sort familiar from physics, in the sense of (minimally) regularities that are invariant under all changes in the values of the relevant variables and parameters. On the other hand, if natural selection is seen to be a process that is itself only locally invariant, then this implies there are more fundamental invariances. These will feature in the relevant physics so that the appropriate laws plus local conditions underpin the range of invariance associated with natural selection. In other words, according to Rosenberg if biological structures are conceived of as spatio-temporally limited and evolving structures, this needs to be understood as holding within a more encompassing or fundamental structure. Then, as Rosenberg indicates in the quote above, there are two options: if biology is not reduced to chemistry and physics, then this more encompassing structure will be that of the principles of natural selection, understood as globally invariant nomological generalizations as in physics; or, if reductionism holds, then this more fundamental structure will be physical structure. So, either we have a *sui generis* form of structural realism for the biological domain, or, ultimately, biological structure is reduced to the kinds of structures presented in and described by theories in physics.

Nevertheless, if we take the first option and attempt to articulate a biological form of structural realism, and even granted that we can substitute models for laws, we do not typically find the other feature of physical structures in biology, namely symmetries. However, one can identify what might be called similarly ‘high-level’ features of biological structures. There is, of course, Price’s Equation (for discussion see Okasha 2006, §1.2 and, in a different context, Rowbottom 2010), sometimes presented as representing ‘The Algebra of Evolution’, and which one could take as characterising a certain fundamental – if, perhaps, abstract – and ‘high-level’ feature of biological structure (French 2011). Put simply, this states that,

$$\Delta \mathrm{z} = \mathrm{Cov}\,(\mathrm{w}, \mathrm{z}) + \mathrm{Ew}\,(\Delta \mathrm{z})$$

where, Δz is the change in average value of character from one generation to next; Cov(w,z) represents the covariance between fitness w and character (action of selection) and Ew(Δz) represents the fitness weighted average of transmission bias (difference between offspring and parents). Thus the equation separates the change in average value of character into two components, one due to the action of selection, and the other due to the difference between offspring and parents. There is a sense in which this offers a kind of ‘meta-model’ that represents the structure of selection in general (for a useful overview, see Gardner 2008; also Okasha 2006, §1.2 and Jones 2008). Okasha writes that it reveals the ‘common logic underlying all selection processes, at all scales and at all hierarchical levels’ (Okasha, op. cit.; see also Okasha 2011) Indeed, as Gardner notes, it can be viewed as reflecting an even more general feature of reality:

> The importance of the Price equation lies in its scope of application. Although it has been introduced using biological terminology, the equation applies to any group of entities that undergoes a transformation. (Gardner op. cit., p. 199)[1]

Although obviously not a symmetry such as those we find in physics, this covariance equation can nevertheless be regarded as describing a high-level feature of biological structure. As Rosales has emphasised, it is independent of objects, rests on no contingent biological assumptions and represents the modal, relational structure of the evolutionary process (see Rosales forthcoming). Just as the laws and symmetries of physics 'encode' the relevant possibilities, so Price's equation encodes how the average values of certain characters changes between generations in a given biological population.

3 One Tool in the Toolbox

Waters, on the other hand, takes the above formulation to represent simply a partial decomposition of evolutionary causes, as he sees the Price equation as just one tool in the toolbox that biologists have available (Waters 2011). Indeed, he urges us to move away from the theory-oriented stance with which structural realism might be seen to be associated. Thus, for example, he conceives of genetics as a science organized by an integration of explanatory reasoning (associated with a theory) and, crucially, *investigative strategies* (Waters 2008). Here the emphasis is on bottom up manipulability and practice and the core feature of the Genetic Approach is that '[g]enes are used as levers to manipulate and investigate a wide variety of biological processes.' This might be regarded as urging a shift back towards an object oriented stance but, as I shall suggest, even this form of 'toolbox realism' can be brought under the structuralist canopy.

So, consider Chakravartty's 'semi-realism' (1998, 2007) that brings together entity realism and structural realism in an attempt to pin down the best of both object-oriented and structuralist stances. Thus the kinds of structures we should be realist about are conceived of in terms of relations holding between first-order, causal properties of objects (2007, p. 41). Those properties that are 'causally linked to the regular behaviours of our detectors' are the 'detection' properties (ibid, p. 47), '... in whose existence one most reasonably believes on the basis of our causal contact with the world.' (ibid), to be distinguished from the 'auxiliary' properties, where the latter have an unknown ontological status, since detection based grounds are insufficient to determine whether they are causal or not. It is in terms of the

[1] As Gardner goes on to note, Price himself emphasized that his equation could be used to describe the selection of radio stations with the turning of a dial just as easily as it could to describe biological evolution.

detection properties that we come to identify the entities that are the focus of the 'entity' realist, and it is these properties that provide the minimal interpretation of the mathematical equations favoured by the structural realist.

Paraphrasing the core slogan of Hacking's entity realism as, 'If you can lever them, they're real', Waters' investigative strategies can be straightforwardly brought within the remit of Chakravartty's semi-realism. But note that this does not immediately give us genes-as-objects, just as Hacking's entity realism does not yield electrons-as-objects. The basis of a well-known criticism of entity realism is that manipulability – spraying in the case of electrons, leverage in the case of genes – is only achieved via certain causal properties, namely Chakravartty's detection properties. These in turn can be understood – according to Chakravartty (2007) – in terms of dispositions for those relations that make up the concrete structures about which we should be (structural) realists. Of course one might argue that we still get genes-as-objects indirectly, as the 'seats' of these causal properties, in terms of which they can be used as levers within the investigative strategies of the genetic approach, but we can begin to see how Waters' toolbox view can be given a structuralist gloss.

Indeed, I think we can go even further and 'de-seat' genes-qua-objects as the locus of these causal properties, leaving an object-empty form of structuralism.

4 Biology without Objects

I have previously noted that issues regarding the role and nature of objects arise in biology also. Of course, these are not the same issues as in quantum physics but they nevertheless motivate a move away from an object-oriented stance (French 2011, forthcoming). These issues include the following: that the notion of 'gene' has undergone such a radical transformation during the history of genetics that there are simply no straightforward identity conditions that it could be said to satisfy throughout the course of that history (Fox Keller 2000); criticism of the 'gene-centred' stance in foundations of biology that has emerged from 'Developmental Systems Theory' (Oyama et al. 2001); the units and levels of selection debate Okasha (2006); the adoption of a 'metagenomic' stance which represents a shift in focus away from individual genomes to 'large amounts' of DNA 'collected from microbial communities in their natural environments' (Dupré and O'Malley 2007, 2009); the heterogeneity of biological entities (Clarke and Okasha 2013; Godfrey-Smith 2011).

As in the case of physics, these issues push us to adopt a structuralist line. According to this, there are no biological objects (as metaphysically substantive entities), all there is, are biological structures, inter-related in various ways and causally informed. Putative objects, such as genes, individual organisms etc. should be seen as dependent upon the appropriate structures ('nodes') and from the realist perspective, eliminable, or, at best, regarded as secondary in ontological priority. This then accommodates the 'fluidity' and 'ephemerality' of biological organisms (as evidenced in the example of symbiotes, for example). Again, from this perspective, biological individuals come to be seen as nothing more than abstractions from the more fundamental biological structure (cf. Dupré and O'Malley 2007), or as 'temporarily stable nexuses in the flow of upward and downward causal interaction' (ibid, p. 842) This still allows for there to be appropriate 'units of selection', but such units are not to be conceptualised in object oriented terms. In particular, we can accommodate the view that, '... a gene is part of the genome that is a target for external (that is, cellular) manipulation of genome behaviour and, at the same time, carries resources through which the genome can influence processes in the cell more broadly.' (ibid).

What is meant by the structure being 'causally informed' here? There are several options. One could follow Dupré and O'Malley and insist that these causal powers are derived from the interactions of individual components and are controlled and coordinated by the causal capacities of the 'metaorganism'. This sort of account seems entirely amenable to a structuralist metaphysics. Alternatively, one could acknowledge that causation is a kind of 'cluster' concept, under whose umbrella we find features such as the transmission of conserved quantities, temporal asymmetry, manipulability, being associated with certain kinds of counterfactuals and so on. Even at the level of the 'everyday' this cluster may start to pull apart under the force of counterexamples. And certainly in scientific domains only certain of these features, at best, apply: thus, understanding causation in terms of the transmission of mass-energy may seem plausible in the context of Newtonian mechanics but it breaks down in General Relativity, where conservation of mass-energy does not apply. Likewise, establishing temporal asymmetry is famously problematic in the context of physics and here we can perhaps, at best, only say that a very 'thin' notion of causation holds, understood in terms of the relevant dependencies. Thus, we may talk, loosely, of one charge 'causing' the acceleration of another charge, but what does all the work in understanding this relationship is the relevant law and from the structuralist perspective, it is that that is metaphysically basic and in terms of which the property of charge must be understood. It is the law – in this case and in the classical context, Coulomb's Law – that encodes the relevant dependencies that appear to hold between the instantiations of the property and that, at the phenomenological level, we loosely refer to as causal.

But once we move beyond physics, the possibility arises of 'thickening' our concept of causation in various ways. We might, for example, insist that for there to be causation there must be, in addition to those conditions corresponding to what are designated the 'cause' and the 'effect', a process connecting these conditions, where this actual process shares those features with the process that would have unfolded under ideal, 'stripped down' circumstances in which nothing else was happening and hence there could be no interference (Hall 2011, p. 115). Such processes can be termed 'mechanisms' (ibid) and here one might draw upon mechanism based accounts of causation and explanation (see, for example, Machamer et al. 2000; for a useful critique, see Psillos 2011). In particular, if such accounts were to drop or downplay any commitment to an object-oriented stance, possible connections can be established with various forms of structuralism.

Thus McKay-Illari and Williamson (2013) have noted that most characterisations of mechanisms can be broken down into two features: one that says something about what the component parts of the mechanism are, and another that says something about the activities of these parts. They advocate an interesting dual ontology with activities as well as entities – of which the parts of mechanisms are composed – in the fundamental base. Here consideration of putative asymmetries between activities and entities (ibid) mirrors to a considerable degree consideration of, again putative, asymmetries between objects and relations within the structuralist context. Indeed, a useful comparison has been drawn (McKay Illari forthcoming) between the insistence that activities are not reducible to entities, so that one needs both in one's ontology and certain forms of 'moderate' structural realism that set objects and relations ontologically on a par (Esfeld and Lam 2008). Thus,

> Activities are real causes, they give us the modal structure of the bundles of mechanism schemas that are biological theories. And biological entities do indeed depend on biological structure. So we have both the basic realist claim, that is also recognizably structural due to a characteristic dependence claim. (McKay Illari forthcoming, p. 12)

However, McKay Illari argues that one can go even further and identify a deeper structure, namely that corresponding to the functional causal roles that are experimentally established in developing mechanism schemas (ibid). On her view, both entities and activities alike should be regarded as the 'locators' of the patterns that Ladyman and Ross focus on in their version of structural realism (2007). The crucial difference between biology and physics, is that in the former '…these patterns are local, patchwork and diverse, which is why we need many many locators to track them' (ibid, p. 16) Both entities and activities can be regarded as locally specific locators for the production of the phenomena that act as explananda and this, she claims, presents us with a 'deep priority of structure', corresponding to that which persists through theory change, yielding a full-blown biological form of OSR. Returning to the issue of biological objects, one could press further and argue that the kinds of examples that are typically given to establish the ontological fundamentality of entities are either 'toy' examples that do not match actual science or simply break down under further examination. Certainly, biological 'entities' seem to be much more fluid and ephemeral than might be initially supposed and this can be taken as motivation for shifting the ontological focus to the relevant activities and processes in the manner that McKay Illari advocates.[2] Precisely how these might be understood from the structuralist perspective still requires further work, of course, but the point is that causality can then be 'de-seated' from objects and possible connections open up with activity-based accounts of biological processes.

Furthermore, to say that genes are eliminable *qua metaphysically robust objects* (in the sense of being the seat of causation, for example) is not to say that we cannot

[2] Alternatively, one could argue that this fluidity supports a view of biological entities as *vague* objects. However, although attempts have been made to articulate a form of ontological vagueness in the quantum context (French and Krause 2003), further work needs to be done to advance this idea within the biological domain.

talk of genes. Eliminativism generates vigorous debate but there are metaphysical techniques we can appeal to in order to mitigate its impact. In this regard metaphysics can be treated as a kind of toolbox from which we can avail ourselves of various strategies, techniques and devices (French and McKenzie 2012). Here's one such technique: according to so-called 'truthmaker theory', the ontological commitments of a theory are not whatever is referred to by the variables of an appropriately regimented theory, as Quinean approaches to ontological commitment insist, but are just those things that have to exist in order to make the relevant sentences of the theory true. Now, on the standard understanding of this account, the truthmaker for the claim 'x exists' is always x and thus in the case of 'Genes exist', we must be committed to the existence of genes. However, one can modify this approach in order to shift ontological commitment elsewhere:

> I think one of the benefits of truthmaker theory is to allow that < x exists > might be made true by something other than x, and hence that '*a* exists' might be true according to some theory without *a* being an ontological commitment of that theory. (Cameron 2008, p. 4)

The core idea here is that what makes the sentence 'Genes exist' true are whatever we take genes to be reducible to or dependent upon. This manouevre allows us to accept that 'Genes exist' is true but refrain from any ontological commitment to genes, because 'Genes exist' is made true by something other than genes-as-objects. Thus, if genes are nothing more than temporarily stable nexuses, then one candidate for what makes such statements true are the relevant features of biological structure. It is to this that we should be ontologically committed, but that doesn't mean we can't talk of genes. Here the gene is seen as a phenomenological entity, not a metaphysically fundamental (at the biological level) object. Alternatively we can allow reductionism to raise its ugly head and insist that what makes sentences such as 'genes exist' true are the physical constituents of genes: molecules, atoms and ultimately elementary particles, which themselves can be reconceptualised in structuralist terms of course.

5 Having the Layer Cake and Eating it

According to Waters, both reductionism and anti-reductionism share a 'layer cake' picture, according to which biology is composed of different layers of organization, distinguished by the different theoretical principles employed[3] (2008). He insists that this is a misleading picture as it encourages philosophers to focus on explanatory theories, rather than research practices, with the latter exemplifying the 'retooling' of classical genetics with genes as levers, as noted above. But as I also argued, this does not imply that genes must be regarded as objects and the Hackingesque emphasis on manipulability can be accommodated in terms

[3] As Waters notes, Weber is an exception.

of the relevant features of biological structures, following the route already laid out by Chakravartty. These structures should themselves be seen as inherently spatio-temporally limited and evolving (this marking a crucial difference from the structures of physics). If natural selection is then taken to represent a global invariance, then following Rosenberg we can think of these structures as again forming layers; but if we take it as local then, as Rosenberg suggests, we must acknowledge the existence of more fundamental invariances.

In this sense, then, a structuralist approach suggests that we can indeed have our cake and eat it – we can acknowledge the presence of explanatory layers whilst also accommodating the manipulability associated with genes as items in the biologist's toolbox. But of course, there is nothing here that compels us to abandon eliminativism: all these tools are ultimately, are fundamental aspects of structure that are 'arranged gene-wise'. This is nothing for the biologist to worry about, since her practices continue unimpeded but the philosopher has a different job of work to do and eliminativism allows her to do it in a clean and simple way.

References

Beatty, J. (1995). The evolutionary contingency thesis. In G. Wolters & J. G. Lennox (Eds.), *Concepts, theories, and rationality in the biological sciences* (pp. 45–81). Pittsburgh: University of Pittsburgh Press.

Cameron, R. (2008). Truthmakers and ontological commitment: Or how to deal with complex objects and mathematical ontology without getting into trouble. *Philosophical Studies, 140*, 1–18.

Chakravartty, A. (1998). Semirealism. *Studies in History and Philosophy of Science, 29*, 391–408.

Chakravartty, A. (2007). *A metaphysics for scientific realism*. Cambridge: Cambridge University Press.

Clarke, E., & Okasha, S. (2013). Species and organisms: what are the problems? In P. Huneman, & F. Bouchard (Eds.), *From groups to individuals: Perspectives on biological associations and emerging individuality* (pp. 55–76). Cambridge: MIT Press.

Dupré, J., & O'Malley, M. (2007). Metagenomics and biological ontology. *Studies in the History and Philosophy of the Biological and Biomedical Sciences, 28*, 834–846.

Dupré, J., & O'Malley, M. (2009). Varieties of living things: Life at the intersection of lineage and metabolism. *Philosophy and Theory in Biology, 1*, 1–25.

Esfeld, M., & Lam, V. (2008). Moderate structural realism about space-time. *Synthese, 160*, 27–46.

Fox Keller, E. (2000). *The century of the gene*. Harvard: Harvard University Press.

French, S. (2011). Shifting to structures in physics and biology: A prophylactic for promiscuous realism. *Studies in History and Philosophy of Biological and Biomedical Sciences, 42*, 164–173.

French, S. (2012). The resilience of laws and the ephemerality of objects: Can a form of structuralism be extended to biology? (pp. 187–200). In D. Dieks, et al. (Eds.), *Probability, laws and structures*. Dordrecht: Springer.

French, S. (forthcoming). The structure of the world: representation and metaphysics. Oxford: Oxford University Press.

French, S., & Krause, D. (2003). Quantum vagueness. *Erkenntnis, 59*, 97–124.

French, S., & McKenzie, K. (2012). Thinking outside the toolbox: Towards a more productive engagement between metaphysics and philosophy of physics. *The European Journal of Analytic Philosophy, 8*, 42–59.

Gardner, A. (2008). The price equation. *Current Biology, 18*, R198–R202.

Godfrey-Smith, P. (2011). *The evolution of the individual*. Lakatos Award Lecture, LSE. (preprint).

Hall, N. (2011). Causation and the sciences. In S. French & J. Saatsi (Eds.), *The continuum companion to the philosophy of science* (pp. 96–119). London: Continuum Press.

Jones, J. H. (2008). Notes on the price equation. http://www.stanford.edu/~jhj1/teachingdocs/Jones-PriceEquation.pdf. Accessed 3 Oct 2013.

Ladyman, J., & Ross, D. (2007). *Everything must go*. Oxford: Oxford University Press.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.

McKay Illari, P. (forthcoming). Is activity-entity dualism ontic structural realism for the biological sciences?

McKay Illari, P., & Williamson, J. (2013). In defence of activities. *Journal for General Philosophy of Science, 44*, 69–83.

Mitchell, S. (2003). *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.

Odenbaugh, J. (2008). Models. In S. Sarkar & A. Plutynski (Eds.), *Blackwell companion to the philosophy of biology*. Oxford: Blackwell.

Okasha, S. (2006). *Evolution and the levels of selection*. Oxford: Oxford University Press.

Okasha, S. (2011). Reply to Sobers and Waters. *Philosophy and Phenomenological Research, 82*, 241–248.

Oyama, S., Griffiths, P. E., & Gray, R. D. (2001). *Cycles of contingency: Developmental systems and evolution*. Cambridge: MIT Press.

Psillos, S. (2011). The idea of mechanism. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 771–788). Oxford: Oxford University Press.

Rosales, A. (forthcoming). The metaphysics of natural selection: A structural approach. Presented at the Annual Conference of the *BSPS* 2007.

Rosenberg, A. (2011). Why do spatiotemporally restricted regularities explain in the social sciences? *The British Journal for the Philosophy of Science, 63*, 1–26.

Rowbottom, D. (2010). Evolutionary epistemology and the aim of science. *Australasian Journal of Philosophy, 88*, 1–17.

Waters, C. K. (2011). Okasha's unintended argument for toolbox theorizing. *Philosophy and Phenomenological Research, 82*, 232–240.

Waters, C. K. (2008). Beyond theoretical reduction and Layer-Cake antireduction: How DNA retooled genetics and transformed biological practice (pp. 238–262). In *The Oxford handbook of philosophy of biology*. Oxford: Oxford University Press.

Must Structural Realism Cover the Special Sciences?

Holger Lyre

Abstract Structural Realism (SR) is typically rated as a moderate realist doctrine about the ultimate entities of nature described by fundamental physics. Whether it must be extended to the higher-level special sciences is not so clear. In this short paper I argue that there is no need to 'structuralize' the special sciences. By mounting concrete examples I show that structural descriptions and structural laws certainly play a role in the special sciences, but that they don't play any exclusive role nor that they give us any reason to believe that all that there is on the various levels is structure. I fortify my points by arguing that structures are global entities (in order for SR not to collapse into a bundle ontology) and that the assumption of higher-level structures as genuinely global or holistic entities is even more arcane.

Many proponents and opponents of structural realism alike seem to agree on the point that SR, if sound, must provide more than just a metaphysics of the fundamental physical level, but that it should also provide an ontological framework that covers higher levels of complexity and thus pertains to the special sciences as well. As Frigg and Votsis (2011, p. 269) put it:

> the question remains whether OSR, and ESR for that matter, can give an adequate account of the ontology and epistemology of other sciences. The bulk of the literature on SR has thus far focussed on modern (and in particular fundamental) physics. This is no accident of history. A structuralist analysis of scientific theories usually departs from those theories' mathematical formalism, and formalisation is the hallmark of modern physics. Therefore SR seems to be at odds with less formal sciences such as biology or the social sciences. This has led some critics to claim that SR is a philosophy with little, if any, relevance outside the province of physics . . .

H. Lyre (✉)
Philosophy Department, University of Magdeburg, Magdeburg, Germany
e-mail: lyre@ovgu.de

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_31,

A few recent authors already took up the challenge and tried to indicate how SR motives can be implemented in the special sciences, as for instance French (2011, 2012) for biology, Kincaid (2008) for social science and Ross (2008) for economy. I consider such attempts as superfluous. In this paper I will argue that there is no need to 'structuralize' the higher-level special sciences. I actually want to show that, quite in contrary, we should not even assume that the special sciences can be provided with the same ontological framework than fundamental physics. In the first section I distinguish between epistemic and ontic SR arguments about the special sciences. I then consider three examples of higher-level structures in the second section. Here my discussion touches on issues of multiple realizability and the possibility of higher-level structural laws. In the third section I argue that OSR is better off to construe structures as global entities to prevent a collapse into a bundle ontology, but that the assumption of higher-level structures as genuinely global or holistic entities is even more arcane. I finish in the fourth section with a short discussion of the possible combinations of SR with scientific (anti-)fundamentalism.

1 ESR and OSR about the Special Sciences

Why should one want to extend structural realist motives to the higher-level, more complex sciences? The answer might depend on whether one adopts an epistemic or ontic point of view. Indeed, the vast majority of the more recent authors in the debate about structural realism focuses on OSR. And rightly so, I think. Structural realism is first and foremost an ontological framework that provides us with a tailor-made metaphysics for modern physics. I will thus adopt an OSR perspective in this paper as well. To start with, however, I shall briefly consider the title question from an ESR point of view. Epistemic structural realists emphasize that our structuralist conception of the world is due to our peculiar and perhaps limited epistemic access to the world. A higher-level ESRist must therefore find arguments why our access to the world on all levels of complexity is restricted to structures rather than object-like entities. The point would go through if, indeed, all of our science were basically formulated in terms of structural laws and regularities with non-individual entities and the like. But just the opposite seems to be, at least mostly, the case (*pace* French, Kincaid and Ross). Quite generally, in our higher-level, special sciences the entities considered on the particular levels (as, for instance, the biological, geological, psychological, sociological, or economic level) are construed as individual entities with intrinsic natures – natures that can be captured by the laws and regularities of the corresponding special science.

Curiously enough, despite *prima facie* evidence, the above mentioned authors not only want to defend ESR but a stronger and more ambitious OSR about the special sciences. Roughly speaking, special science OSR is the view that all that there *is* on the various levels of complexity or description on which the special sciences operate is structure. A motive for a proponent of OSR to extend her view to the special sciences is scientific anti-fundamentalism – the belief that there is no bottom level. This attitude is sometimes captured by the slogan that it is "structures all the way down". While I think that physics gives us strong reasons to believe in the existence of a bottom level rather than believing in scientific anti-fundamentalism, I don't delve into arguments for this claim here. What I want to point out is that scientific anti-fundamentalism together with genuine higher-level OSR commits the proponent of such a view to the existence of 'genuine' structures on *all* levels. No level is fundamental, no level will serve as the bottom reduction or supervenience base. All levels are, as it were, genuine and consist of *bona fide* structures. And, of course, it's not only structures all the way down, it's also structures all the way up! Let's see whether the assumptions of such a view are tenable.

2 Three Examples of Higher-Level Structures

From now on I focus on OSR, but my point in this section, if successful, can be understood such that even special science ESR is in bad shape (and within the context of our discussion undermining ESR *a fortiori* undermines OSR as the more ambitious position). My point is the following: while I do believe that structural descriptions and structural laws play an eminent role in the special sciences, I fail to see that they play an *exclusive* role or that they should give us any reason to believe that all that there *is* on the various levels is structure. SR proponents don't need to assume the existence of genuine level-bound structures since, in general, higher-level structures simply supervene on or can otherwise be traced back to lower-level features. I shall present three examples to illustrate this.

Consider, as a first example, the harmonic oscillator. It provides us with a simple and straightforward example of a system that is described by purely structural means. At the same time, and exactly for the same reason, it provides us with a perfect example of multirealizability: harmonic oscillators are instantiated by pendula, springs, electromagnetic circuits, neural circuits etc. All harmonic oscillators are governed by the same structural law: the oscillator equation $\mathrm{d}^2/\mathrm{dt}^2\ \mathrm{x(t)} + \mathrm{k\ x(t)} = 0$. It is a structural law in the sense that, as far as the oscillation is concerned, no intrinsic but only relational properties of the target system play a role. That's why harmonic oscillators are multirealizable: they all share the same (sub-) set of relational properties (obeying the same regularity). The various classes of instances of the harmonic oscillator are individuated by the constant *k* only (*k* defines what *kind* of oscillator we are dealing with: spring, pendulum, neural circuit etc.). The example of the harmonic oscillator expedites an important insight about the nature of multirealizability. As I have argued elsewhere (Lyre 2009), a majority of cases of multirealizability simply rests on the sharing of properties being either intrinsic or even purely relational properties. While cases of shared intrinsic properties are non-thrilling cases of multirealizability (e.g. the property of "being red" is trivially multirealized), harmonic oscillators are of the more interesting type of cases of shared relational properties.

This directly relates to laws, since laws quantify over shared properties. Take "like charges repel" as a simple example of a law. Like charged particles repel irrespective of their mass, size and spin. So laws quantify over particular properties and disregard others. In this brute sense, laws and in particular higher-level laws are "multirealizable". But these are benign cases of multirealizability of course, since the causally relevant properties are just shared lower-level properties. And this is also true in the case of structural laws such as the oscillator equation (and, as I've argued in my 2009, even in the case of functional laws), where the causally relevant properties are shared lower-level *relational* properties. But what is most important for the purpose of our discussion is that structural laws, as far as they occur in the special sciences, can be reduced to lower-levels insofar as the purportedly existing higher-level structures simply supervene on lower-level ones (i.e. sets of shared lower-level relational properties).

So let's go over to a second example. I do not want to claim that higher-level structures can in all cases be (Nagel-) reduced to lower-level ones. Surely in many important cases of higher-level structural laws such laws also include epistemically advantageous and tricky approximations and idealizations. Here's a simple toy example for the "addition rules of huge numbers". Consider simple addition cases like $12 + 36 = 48$ and $3 + 3 = 6 = 2 \cdot 3$. The general rules can be captured as $a + b = c$, where $c > a,\ b$ and $a + a = 2a$ with $a,\ b,\ c$ being 'small' natural numbers. Now consider the addition of huge numbers like $10^{80} + 10^{120} \approx 10^{120}$ and $10^{80} + 10^{80} = 2 \cdot 10^{80} \approx 10^{80}$. For all practical purposes, the rules for the addition of huge numbers can be generalized as $A + B = B$, where $B >> A$ and $A + A = A$, where $A,\ B$ are 'huge' numbers (I don't think it's necessary for the purpose of my illustration to define 'small' and 'huge' more rigorously here). What we can see from this simple example is how the higher-level addition structure "emerges", as it were, from the lower-level addition structure of normal and small natural numbers.

My second example can be supplemented by a third one, the example of John Conway's infamous "Game of Life", the probably best-known example of a cellular automaton. 'Life' is fascinating since it opens up a universe of complexity and a plethora of patterns that live on an infinite two-dimensional grid of cells while being based on just three amazingly simple rules. These so-called updating rules for the state of a cell as being either alive or dead simply are: (1) live cells with $\mathrm{n} < 2$ die (by loneliness), (2) live cells with $\mathrm{n} > 3$ die (by overcrowding), and (3) dead cells with $\mathrm{n} = 3$ come to life (n is the number of neighbours). The cell patterns that arise during the temporal evolution of a particular game starting from a particular initial state can be seen as higher-level structures that all and only consist of lower-level pixel distributions.

The question arises whether there's ontologically more to such patterns than their undeniable epistemic and pragmatic value from an instrumentalists's and interpretationalist's point of view. Daniel Dennett calls them 'real patterns' (Dennett 1991), but remains vague about the strength of his ontological commitment to them. Ladyman and Ross are more explicit insofar as they subscribe to a *scale relativity* *of ontology* understood as the thesis that *"what (really, mind-independently) exists should be relativized to (real, mind-independent) scales at which nature is measurable"* (Ladyman and Ross 2007, p. 200). However, they go on to claim that *"because in Life there is an unambiguous fundamental level composed of the aggregation of a finite number of 'little things', and because no higher-level object types cross-classify the dimensions of any models of the game relative to classifications in terms of cells, Life differs greatly from the universe with respect to the kinds of reductionism sustainable in it. Life admits of complete decomposition; the universe does not"* (Ladyman and Ross 2007, p. 201, fn. 12). It remains of course their task to prove the latter anti-reductionist claim.[1]

For our purposes, the lesson from the three examples is that structural laws may occur on all levels of complexity and in all domains of science, but that it is very reasonable to assume that they either rest upon shared relational properties (as in the example of the harmonic oscillator) or on approximations or idealizations of lower-level properties (as in the second and third example). Under these assumptions, however, higher-level structures aren't *bona fide* higher-level entities. They are not what a higher-level structuralist must expect them to be.

3 More Evidence Against Higher-Level OSR: Structures as Global Entities

OSR must be distinguished from a bundle ontology, where objects are construed as property bundles (non-individual objects may thus be construed as bundles of relational properties). But bundles are still atomistic or pointillistic entities, they are considered to exist at spacetime points. OSR, however, should be conceived as holistic. As I've argued in my 2012, structures aren't mere collections of local relational properties, but *global* entities reflecting features of the world in toto. Otherwise, OSR collapses to a bundle ontology. At the same time, global structures are *in re*-structures in the sense that they only exist insofar as they are causally efficacious. For instance, the Lorentz structure of Minkowksi spacetime is causally efficacious insofar as it endows spacetime with a geochronometrical structure of inertial trajectories. Or take the U(1) gauge group. It exists only insofar as it deploys actually occurring causal effects. They can finally be detected in all of our experiments in connection with electromagnetic interactions.

[1] A remark about the notorious talk of "cross-classification" (cf. my 2009): the widespread anti-reductionist claim that higher order properties cross-classify lower-level ones, can, as Kim (1998, p.69) has rightly pointed out, only reasonably be maintained if one is willing to give up supervenience. For two taxonomies to cross-classify opens the possibility that the higher-level taxonomic class makes causally efficatious distinctions not made by the lower-level one. But this is a clear failure of supervenience. Cross-classifying taxonomies define conflicting causal profiles.

At least two clarifications are in order here. First, note that the structuralist doesn't want to say that there exist charged particles with interactions but rather that the observations we do in our labs and experiments, and that we superficially attribute to particles, must be traced back to a global U(1) quantum gauge structure of the world that is causally efficacious and, hence, brings about such observable effects. Another way of putting it would be to say that those effects are instantiations or realizations of the global structure. I don't see a particular problem with this phrasing except that many read this as a Platonistic statement: there's an abstract structure on the one hand and it's world-like realization or instantiation on the other. But this, of course, were to confuse *in re-* with *ante rem*-structuralism, which is neither intended nor enforced by any of the above. As being *in re* and concrete, structures should not be confused with abstract Platonistic entities. They are global and concrete rather than local and abstract.

A second point of clarification has to do with "causal efficacy". For a modal structuralist as Esfeld (2012) a structure *necessarily* brings about its observable effects. For a Humean structuralist as myself the structure just brings about certain effects. They are whatever we observe them to be. No necessity is involved. For both of us, I take it, the structure only exists insofar as its causal efficacy is actualized – meaning that even Esfeld doesn't want to say that there are unactualized dispositional structures. Or does he? Anyway, the Humean point of view give us a straightforward and unambiguous picture: structures just are. We know about their existence because of the observations we do in our labs. These observations can be best explained in terms of a scientific realism of a structuralist kind. No necessity is involved. As being global entities, however, we understand that the observations we do in all of our labs and that are spread over spacetime are orchestrated in such a way that they are the offspring of a global structure and not of local goings-on that must, in a second step, somehow be glued to each other by mysterious modal laws.

Let me come back to my main line of thought. That structures should be considered as global structures provides us with an even stronger argument against special science OSR. It is just highly implausible to assume that higher-level structures reflect genuinely higher-level holistic or global world features. At least, I've never seen arguments in favour of such a view. And notice that a structuralist about higher-level science must show that all levels and, accordingly, all special sciences must be interpreted like this: all levels must then consist of all and only structures and, in order not to collapse structuralism into a bundle view, such structures must be considered as global and holistic structures. Why should anybody subscribe to such an arcane view? Even the proponent of Dennettian real patterns of the Ladyman-Ross-style doesn't commit himself to such patterns as global entities. He rather considers them as patterns that are composed of local objects and their relations (think of the game of life patterns and their composite structure). On higher levels of complexity, our localistic picture of the world (as consisting, for instance of chemical molecules, biological organisms, social groups etc.) is an approximation of a world that supervenes on lower levels, perhaps even on some bottom level.

4 Scientific Fundamentalism and OSR

This finally brings me to the issue of scientific fundamentalism. As I've said in the beginning, a motive for the proponent of OSR to extend his view to the special sciences is scientific anti-fundamentalism. But we've seen arguments against the plausibility of special science OSR. How do things stand if we adopt an OSR perspective together with scientific fundamentalism (in short: F-OSR)?

There are in principle two possibilities here. One might be an F-OSRist and a reductionist and end up with the view that all that there is elementary structure. Basically, that's my favourite view. One might, however, also be an F-OSRist and a non-reductionist. Then there are again two possibilities. One might think that only the bottom level consists of structures while all the higher levels consist of object-like entities (or a mixture of object-like entities and structures). In a sense that sounds like a strange and hybrid position. Perhaps a proponent of such a view believes that there exist higher-level object-like entities with intrinsic natures and that it is generally impossible to reduce intrinsic to relational properties. But that is obviously wrong, intrinsic properties may very well supervene on relational properties. Here's a simple example: the property of "being a graph" can be considered an intrinsic property of the whole graph, which, itself, is a purely structural entity.[2] In any case, such a position is not in my present scope since we are interested in the question whether there's any need for higher-level structures.

So, as a second possibility, one might think that both the bottom and all of the higher levels consist of structures. But in that case my preceding objections already apply. And in a sense they apply even stronger. While the anti-fundamentalist will typically also be an anti-reductionist, the F-OSRist has now, in the light of the preceding objections, good reasons to give up anti-reductionism. So, again, we are better off with reductionist F-OSR.

[2] Anyway, OSR is more than the idea that there are "just relations". As I've argued elsewhere (Lyre 2010, 2012), OSR must be supplemented with a weak and special type of intrinsicality. I've dubbed this 'Extended OSR' (ExtOSR) – the view that relational and structurally derived intrinsic properties exist. Simple OSR, by contrast, assumes only relational but no intrinsic properties (however, both ExtOSR and SimpOSR are versions of non-eliminative OSR). Reasons to prefer ExtOSR over SimpOSR are symmetry invariants and zero-value properties (cf. my 2012). Pick up the first reason: The symmetry invariants under a given symmetry over a domain D provide properties that are shared by all members of D. They are 'intrinsic' in the sense that they belong to all members of D irrespective of the existence of other objects. Since they are shared by all members of D, they serve to individuate domains, not individuals. Such structure invariants provide structurally derived intrinsic properties. SimpOSR denies intrinsicality, but symmetry groups come equipped with their invariants. So SimpOSR doesn't have the resources to embrace the symmetry structures of modern physics represented by the fundamental symmetry groups. Moreover, almost all fundamental symmetries are quantum gauge symmetries. Here, the argument about symmetry invariants becomes even more pressing since gauge symmetry transformations possess no real instantiations. Only the gauge invariants do. Hence, ExtOSR must be favoured.

5 Conclusion

I've granted the occurrence of higher-level structural laws, but at the same time I've argued that such higher-level structural laws may typically be understood as supervening on lower-level properties or structures. Nothing commits us to the existence of genuine, non-reducible higher-level structures. I've fortified my points by arguing that structures are global entities and that the assumption of higher-level structures as genuinely global or holistic entities is even more arcane. It is only on the bottom level, where the global and holistic nature of the fundamental structures becomes apparent. It is, accordingly, only reductionist F-OSR that provides us with a genuine structuralism that doesn't collapse to a bundle view. There's all in all no need to structuralize the special sciences.

References

Dennett, D. (1991). Real patterns. *Journal of Philosophy, 88*(1), 27–51.

Esfeld, M. (2012). Causal realism. In D. Dieks et al. (Eds.), *Probabilities, laws, and structures* (pp. 157–168). Dordrecht: Springer.

French, S. (2011). Shifting to structures in biology and beyond: A prophylactic for promiscuous realism. *Studies in History and Philosophy of Biological and Biomedical Sciences, 42*, 164–173.

French, S. (2012). The resilience of laws and the ephemerality of objects: Can a form of structuralism be extended to Biology? In D. Dieks et al. (Eds.), *Probabilities, laws, and structures* (pp. 187–199). Dordrecht: Springer.

Frigg, R., & Votsis, I. (2011). Everything you always wanted to know about structural realism but were afraid to ask. *European Journal for Philosophy of Science, 1*(2), 227–276.

Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press.

Kincaid, H. (2008). Structural realism and the social sciences. *Philosophy of Science, 75*, 720–731.

Ladyman, J., & Ross, D. (2007). *Everything must go: Metaphysics naturalized*. Oxford: Oxford University Press.

Lyre, H. (2009). The "multirealization" of multiple realizability. In A. Hieke & H. Leitgeb (Eds.), *Reduction – Abstraction – Analysis* (pp. 79–94). Frankfurt: Ontos.

Lyre, H. (2010). Humean perspectives on structural realism. In F. Stadler (Ed.), *The present situation in the philosophy of science* (pp. 381–397). Dordrecht: Springer.

Lyre, H. (2012). Structural invariants, structural kinds, structural laws. In D. Dieks et al. (Eds.), *Probabilities, laws, and structures* (pp. 179–191). Dordrecht: Springer.

Ross, D. (2008). Ontic structural realism and economics. *Philosophy of Science, 75*, 731–741.

Part XI
Philosophy of the Cognitive Sciences

Principles Versus Mechanisms in Cognitive Science

Lilia Gurova

Abstract The view that mechanistic explanations best characterize "the explanatory project of cognitive science" Bechtel (*Topics in Cognitive Science 2(3)*, 357–366, 2010) has recently been promoted. The proponents of this view insist that law-like statements in cognitive science could not play any explanatory role because they are mere descriptions of the empirical effects which have to be explained. The aim of this paper is to demonstrate that mechanistic explanations are not "the only game in town" in cognitive science. Principle-based explanations have sometimes been advanced to cope with important empirical findings and the principles involved in such explanations are more than mere descriptions of empirical effects. The role of principle-based explanations is illustrated by the example of basic level effects, one of the most important discoveries about categorization in the last 50 years. The example of basic level effects suggests that under certain conditions the principle-based explanations seem to be the only available choice.

The view that mechanistic explanations best characterize "the explanatory project of cognitive science" (Bechtel 2010, p. 365) has recently been advocated in a series of publications.[1] In support of their view, the proponents of the mechanistic explanatory project have argued that:

(i) Explanations that appeal to laws or any law-like statements are very rare in life sciences in general and in the cognitive sciences in particular. In this

[1] Cummins (2000), Bechtel and Abrahamsen (2005), Wright and Bechtel (2007), Bechtel (2009, 2010, 2011).

L. Gurova (✉)
Department of Cognitive Science and Psychology, New Bulgarian University,
21 Montevideo Str., 1618 Sofia, Bulgaria
e-mail: lgurova@nbu.bg

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_32,

respect, life sciences are essentially different from physical sciences: principles and laws are powerful explanatory tools in physics but they play a rather insignificant role in biology.[2]

(ii) General statements in psychology and other cognitive sciences that look like laws and sometimes are even called so, are mere descriptions of established empirical effects and as such they are not explanatory in respect to these effects.[3]

(iii) Those who insist that laws and law-like statements play an important explanatory role in cognitive sciences are either still under the sway of the deductive nomological (DN) model of scientific explanation or they are not sufficiently familiar with the real practice of cognitive scientists.[4]

In this paper I draw attention to some counterexamples to the claims (i)–(iii). Then I present a case study to illustrate the role of principle-based explanations in cognitive science. The case study is about the so-called basic level effects which, according to Murphy (2002), are one of the genuine discoveries about categorization made in the last 50 years. There have been two major attempts to explain the basic level effects and both of them have ended with principle-based explanations. No suggestions for mechanistic explanations of these effects have been made so far. I argue that this is hardly incidental: the basic level effects constitute a good example of a case where principle-based explanations seem indispensable.

1 Law-Like Statements in Life Sciences: Explanations that Appeal to them are Neither Rare Nor is their Role Insignificant

Some famous examples directly contradict the claim that explanatory appeals to law-like statements in life sciences are rare and/or their role is insignificant.

The most salient example is the principle of natural selection. In its original Darwinian formulation, the principle states that "each slight variation, if useful, is preserved" (Darwin 1859, p. 61). Evolutionary biologists and philosophers of biology still argue whether this statement exemplifies a genuine scientific law, and how exactly it is involved in the explanatory practice of evolutionary biology.[5] Despite

[2] Bechtel (2009, 2010, 2011).

[3] This claim was first stated explicitly by Cummins (2000) but it has been popularized since then mainly by Bechtel – see (Bechtel and Abrahamsen 2005; Bechtel 2009, 2010).

[4] See e.g. Wright and Bechtel (2007), p. 31.

[5] A prominent defender of the view that the principle of natural selection is a genuine scientific principle which is not significantly different from similar principles in physics (like the second law of thermodynamics) is Rosenberg – see (Rosenberg 2001; Bouchard and Rosenberg 2004). Among those who have argued that the law-like statements in biology do not bear the essential characteristics of physical laws are Mayr (1985) and Beatty (1995). Sober, who admits that

these controversies, however, very few respectable scholars have ever questioned the explanatory role of the Darwinian principle.[6] A large amount of interesting phenomena received explanations due to the presumption that useful (adaptive) inheritable variations are preserved. Let's mention only Fisher's explanation of the sex ratio (Fisher 1930), Medawar's explanation of aging (Medawar 1952), or the explanation of microorganisms' resistance to antibiotics.[7]

Another famous law-like statement in life sciences is the so-called "central dogma of molecular biology" which states that "once (sequential) information has passed into protein it cannot get out again" (Crick 1958, p. 153). The central dogma was introduced in 1958, in a situation which Crick himself described later as "a period when much of what we now know in molecular genetics was not established" (Crick 1970, p. 561). In this situation of "fragmentary" and "rather uncertain and confused" experimental results, the central dogma was expected to help "stating problems clearly and thus guiding experiments" (Crick 1970, p. 561). Fifty years after the launch of the central dogma, biologists and philosophers of biology disagree about its proper explanatory role. Some claim that it "has proved to have extraordinary explanatory power" (Botstein 1995, p. 3), others assess its role as rather modest (Sarkar 2005). This controversy, however, does not change the fact that together with the principle of natural selection, the central dogma "is believed to provide the underpinning to all of biology" (Morange 2009, p. 236).

The list of explanatory principles in life sciences is not exhausted by the central dogma and the principle of natural selection. We can add in Mendel's laws of inheritance, Hardy-Weinberg law about the constancy of genotype frequencies in large populations, Kleiber's law about the connection between the energy consumed by an organism and its weight, and many others. In the face of the impressive history of explanatory success of these principles, it is hardly reasonable to keep arguing that (i) that explanations that appeal to law-like statements are rare and/or play an insignificant role in life sciences.

Here is the place to bring into focus a long discussed issue about whether the above-mentioned biological principles are contingent or not, or whether they hold universally across space-time or not. What I am arguing for here is that the questions about the contingency and the universality of the biological principles are beside the point when we are interested in whether these principles have actually been used in biological practice to explain observed phenomena. The principles discussed in this section have been introduced, and successfully used, for explanatory reasons.

referring to laws is rare in biological explanatory practice, has argued at the same time that models in biology play essentially the same role as laws in physics: "they are general, they don't refer to specific places, times, or individuals, and they support counterfactuals" (Sober 2008, pp. 45–46).

[6] Fodor and Piatelli-Palmarini are famous exceptions: (Fodor 2008; Fodor and Piatelli-Palmarini 2010).

[7] Many of these explanations have led to the formulation of new testable hypotheses which have been submitted to empirical tests and confirmed.

2 Law-Like Statements in Cognitive Sciences: Many of them are More than Mere Descriptions of Empirical Effects

Philosophical skepticism about the possibility of psychological laws has a long history which began with Kant's *Critique of Pure Reason* where Kant stated that psychology would never become a true science because of the lack of good candidates for (a priori true) psychological laws. The subsequent history confuted Kant's pessimism: the discovery of the Weber-Fechner law traced the road to the rise of scientific psychology. Contemporary philosophical skepticism about psychological laws takes a different stance. Those who express such skepticism insist that the alleged laws in psychology are not the same as the laws in natural sciences (e.g. the laws of physics) and that, respectively, their role is different. The candidates for psychological laws have mostly been blamed for being "supple", "soft", or only "ceteris paribus" valid. And recently, a new conviction has been added to the list: that all law-like statements in psychology are descriptions of empirical effects and as such they are not explanatory in respect to the effects which they describe. In order to show why this criticism is difficult to sustain, let's first look at a classical physical law: Newton's law of universal gravitation. This law states that two physical bodies attract each other with a force which is proportional to the product of their masses and inversely proportional to the square of the distance between them. Does Newton's law of gravitation describe an effect? Yes, we may say that it describes the effect of mutual attraction of physical bodies possessing particular masses. Is that law nevertheless explanatory? Yes, it is. It explains, for example, why a human being could not walk without using special technical facilities, on the surface of a planet which has a mass e.g. five times bigger than the mass of the Earth. Let's now look at a classical psychological law, the Gestalt law of closure (Gesetz der Geschlossenheit) which states that objects that seem to form a meaningful image are seen as a whole (Katz 1950). Does the law of closure describe an empirical effect? Yes, it describes the effect of grouping together otherwise non-connected patterns when the result of the grouping is a meaningful image. Is that law explanatory? Yes, it is. It explains, for example, various perceptual phenomena as ignoring the gaps and "seeing" missing contours in cases such as the one shown in Fig. 1. There are no circles and squares there, but if asked, most subjects would report that they see four dark circles and a white square partially overlapping them.

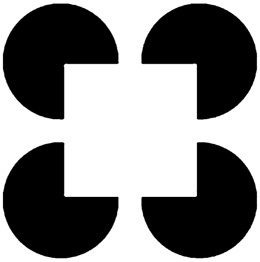


Fig. 1 An illustration of the Gestalt law of closure

Again, it is not relevant to ask here whether the alleged Gestalt law is contingent and whether it holds universally. A law-based inference can be explanatory even if it is not valid in all possible worlds.

It is true, however, that one of the effects of the cognitive revolution which gave birth to the interdisciplinary project of cognitive science was the shift of attention away from the search for psychological laws (Chater and Brown 2008). Chater and Brown explained that by the influence of computer science inside "the new science of mind":

> By viewing the mind as a highly complex computational device, it becomes natural to think of cognitive science as a process of 'reverse engineering' ... rather than following in the mould of physics. Computer science does not seem to be full of quantitative universal laws – instead, its focus is on representations and algorithms operating over those representations. (Chater and Brown 2008, p. 37).

This citation reveals where the temptation of looking for mechanistic explanations comes from. In the same paper Chater and Brown bring into focus the question about the importance of general principles for cognitive science. For them, it is not only possible, but very desirable to arrive at such principles insofar as they "may serve as crucial building blocks for the construction of cognitive theories in specific domains" (Chater and Brown 2008, p. 37). The latter claim is illustrated by the example of Shepard's Universal Law of Generalization for psychological science (Shepard 1987) and by two principles suggested by Chater and Brown themselves[8] that build on Shepard's law. Chater and Brown's position about the place of principles in the explanatory practice of cognitive science is summarized in the following citation:

> It seems entirely possible, and indeed highly likely, that there are many aspects of cognition that must be understood in terms of specific representations and algorithms, which will not be neatly described by universal principles. But each individual case should, we suggest, be considered on its merits – and the possibility that general principles may combined to explain apparently complex phenomena should not be discounted. (Chater and Brown 2008, p. 57).

Chater and Brown's defense of principle-based explanations in cognitive science is not a lone voice in the wilderness. In a recent discussion about the complex systems approach to cognitive science Stephen and Van Orden expressed their worry that

> empirical cognitive science may begin to resemble a parody of reductionism because mechanisms accrue pretty much one to one with empirical effects – as though each empirical effect is visibly transparent to its underlying mechanism – a problem that has been called the *effect* = *structure fallacy* (Gibbs 1994; Lakoff 1987) and the *module mistake* (Van Orden and Kloos 2003; Stephen and Van Orden 2012,p. 95).

[8] Chater and Brown call them respectively "the simplicity principle" and "the scale invariance principle".

In the next section I present a case study that is in tune with Chater and Brown's as well as with Stephen and Van Orden's observations and which suggests that in the case of complex cognitive phenomena, principle-based explanations seem indispensable.

Before going to the case study, however, let's consider the alleged connection between the appreciation of the explanatory role of law-like statements and the adherence to the deductive-nomological model of scientific explanation.

The deductive-nomological model of scientific explanation (shortly, the DN model) is a philosophical invention. Its critics correctly insist that it covers only part of the explanatory inferential relations. For example, the use of the principle of natural selection to explain the preserved phenotypic traits could not be subsumed under the DN model because from "If a variation is useful then it is preserved" (the Darwinian version of the principle of natural selection) and the fact that a particular variation X has been preserved, one cannot deduce that the variation X is useful and thus she cannot argue that X has been preserved because it is useful. But many times evolutionary biologists do arrive at such explanations: that the trait X has been preserved because it is useful. Obviously, the explanatory schema which they follow is not that of the DN model. Similar examples demonstrate that the explanatory use of other biological and psychological laws could not be subsumed to the DN model, too. In the face of such examples it does not make much sense to insist that (iii) those who adhere to principle-based explanations in life sciences in general and in cognitive science in particular do that because of the influence of the DN model. In addition, none of the cognitive scientists who pleaded for the recognition of the principle-based explanation in cognitive science have ever referred to the DN model, either explicitly or implicitly.

3 Principle-Based Explanations in Cognitive Science: The Case of the Basic Level Effects

Brown (1958) was probably the first to draw attention to the fact that people prefer a particular level of categorization in speech: e.g., seeing a dog, most people usually call it "dog" instead of "bulldog" or "animal". About 15 years later Berlin (1972) added a further observation to this one: that in the case of living things (plants and animals) the preferred level of categorization is the same in different cultures. Berlin supposed that this happens because the categories at the preferred level correspond to "natural groups of organisms". Rosch, however, disagreed with this explanation. Her own cross-cultural studies of color categorization revealed that even if there are no natural groupings (the color spectrum is physically continuous) the representatives of different cultures tend to form the same color categories.[9] She viewed that as a crucial support for the claim which she advanced later that there

[9] See Rosch (1973).

must be psychological (in addition to physical) determinants of categorization in all cases, not just in the case of categorization of colors (Rosch 1978).

Rosch made her own contribution to the investigation of the preferred level of categorization which she called "basic level". She and her collaborators found that the members of the basic level categories share a significant number of common attributes, including a similar shape. They found also that the subjects dealing with basic level categories easily form an average image of the category and tend to use this image in categorization tasks, e.g. in tasks where they have to decide whether a particular entity belongs or does not belong to a given category. Rosch et al. found as well that subjects tend to use the same motor programs to deal with the members of the basic level categories.[10] All these findings are known today as "basic level effects". In the late 70s the significance of the basic level effects had been already recognized despite the controversies surrounding some experimental results. It was more or less reliably established that the basic level effects penetrate the whole cognitive system[11] and that most of them are universal across cultures.

In 2002, in a book which had the ambition to summarize the most important empirical findings in the field of categorization, Murphy admitted that basic level effects do constitute a "genuine discovery" but unfortunately, they are not explained by the current theories of categorization. By "current theories of categorization" Murphy designated all the views which (a) share the common assumption that to recognize a set of objects as a category means to have a unique representation for this set of objects; and (b) differ on what is the structure of the alleged category representations. These views do form the mainstream approach to categorization in contemporary cognitive science.[12] In the late 70s, however, Rosch launched a different approach to categorization. Instead of trying to explain how people categorize the world by asking what kind of mental representations they form and use in categorization tasks, she asked what kind of principles rule the process of categorization. Rosch formulated two general principles in order to explain a large set of seemingly unconnected phenomena including the basic level effects. The two principles of Rosch's theory of categorization – "the principle of cognitive economy" (R1) and "the principle of perceived world structure" (R2) – state that:

> (R1): "the task of category systems is to provide maximum information with the least cognitive effort".
>
> (R2): "the perceived world comes as structured information" (Rosch 1978, p. 190).

[10] E.g. when asked to describe the series of movements which they make when use a *chair* (a basic level category), the subjects describe quite similar consequences of actions; however, their reports significantly differ when asked to tell what they typically do with a piece of *furniture* (a category above the basic level) (Rosch 1978).

[11] Manifestations of basic level effects have been found in perception, imagery, motor reactions, language, reasoning, cognitive development. For a summary of these effects see (Rosch 1978; Murphy 2002.

[12] An early and still very influential review of the main views on categorization (the so-called "classical view", the prototype view and the exemplar view) is (Smith and Medin 1981). Later on a new "theory view" was added to the list (Murphy and Medin 1985).

How did Rosch explain the basic level effects by means of these two principles? First, she assumed that the cognitive task postulated by the principle of cognitive economy is achieved when "categories map the perceived word structure". She also assumed that the categories belonging to different levels in a given taxonomy do not map this structure equally well. If this is the case, then the categories belonging to one particular level should best map the perceived world structure. And, according to Rosch, this particular level is precisely the empirically established basic level.

Rosch's principles of categorization are not popular today. They are not, for example, discussed in contemporary textbooks along with the theories that form the mainstream approach to categorization. There is no trivial answer to the question why this is the case. Some reasons, however, suggest that it might be due to the dominance of the "reverse engineering" thinking which Chater and Brown (2008) wrote about and which, they insisted, is a result of the invasion of computer science conceptions and methods in cognitive studies. The problem is that the mainstream views of categorization are easily implemented in computer models while Rosch's principles are not. Whatever the reasons for Rosch's principles of categorization to be disregarded, it is important to stress the following about them and their rivals:

(1) Rosch's principle-based theory is the only one in the field which has provided an explanation of basic level effects;
(2) Rosch's theory of categorization was not *post factum* and ad hoc; in fact most of the discoveries of new manifestations of basic level effects reported by Rosch (1978) were predicted by her theory;
(3) The mainstream mechanistic theories of categorization do not explain basic level effects.

In the 70s, however, a problem about basic level effects was recognized which Rosch's theory could not explain. In the case of folkbiological categorization, the experimental results obtained by Rosch et al. differed from the observations of ethnobiologists: the psychologists identified as "basic" a level which was above the "basic level" observed by ethnobiologists. For example, whereas psychological experiments revealed the categories "tree", "fish" etc. as basic level categories, the observations of ethnobiologists associated the basic level effects with the categories "maple", "salmon" etc. which have a lower taxonomic rank. Rosch guessed that it was the lack of sufficient accuracy of ethnobiological methods that caused the difference. Most authors who discussed this discrepancy after Rosch tended to see it as a manifestation of a context effect due to the different level of expertise demonstrated by urban dwellers (the typical participants in psychological experiments) and by people who live in a more natural environment (the typical subjects of ethnobiological studies). However, the assumption that basic level effects are easily movable by the context seems to contradict the earlier finding that these effects are universal across culture. This seeming paradox was resolved by Medin and Atran (2004) who recently launched another principle-based explanation of basic level effects.

Medin and Atran's theory has a narrower scope than Rosch's theory: it is about folkbiological categorization only. The principles which Medin and Atran's theory is built on are the following:

(MA1): The taxonomic structure of folkbiological classifications is universal across culture and it has three levels: folk generic, life form, and folk kingdom.
(MA2): There is a privileged level of categorization of living things which is guided by the notion of essence: its output coincides with the groups of organisms that are believed to share common essential properties.

Following these two principles, Medin and Atran inferred that the basic level of categorization in a given folkbiological taxonomy is the level at which the categories coincide with the groups that (people believe) share common essential properties.

The crucial difference between Rosch's account of basic level categories and that of Medin and Atran is related to the question what determines the basic level in a given taxonomy. For Rosch, it is the "perceived world structure" that determines the process of categorization. According to Medin and Atran, the leading psychological determinant in categorization is the notion of essence. The notion of essence, they claim, does not depend on perceptual experience, that's why any defects in this experience (e.g. the lack of exposure to natural environment) cannot cause a shift of the basic level and hence it is the same across different cultures having different perceptual experience. But some of the manifestations of the basic level (not the level itself), Medin and Atran admit, could change if the task in which we observe them depends on perceptual experience. This happens, for example, in recognition tasks. Shown the same object, subjects who have been normally exposed to the biological world will recognize it as a "maple" while urban dwellers having limited contact with the natural flora will categorize it as a "tree". Given a reasoning task (e.g. category-based induction), however, both groups of subjects demonstrate the same basic level effects because reasoning based on the notion of essence does not crucially depend on perceptual experience.

Like Rosch's theory of categorization, Medin and Atran's theory of folkbiological categorization is a genuine scientific theory which does not only explain facts that have already been established. This theory has been used to make predictions like the one that the basic level effects manifested in category-based inductive reasoning problems will not depend on subjects' previous perceptual experience. And these predictions were experimentally confirmed.

4 Conclusions

I tried to demonstrate that some recent attempts to justify the claim that mechanistic explanations best characterize "the explanatory project of cognitive science" rely on a distorted picture of the role which principle-based explanations play in cognitive science. The picture presented by the claims (i)–(iii) is distorted because:

(1) As the examples discussed in Sect. 1 reveal, appeals to principles in life sciences (like the principle of natural selection, the central dogma of molecular biology etc.) are neither rare nor insignificant.

(2) The principles that one finds in psychology and cognitive science are more than descriptions of effects and, as the comparison between the law of gravitation and the Gestalt law of closure reveals (Sect. 2), they do not differ significantly from physical principles in this respect.

(3) Many principle-based explanations in life sciences could not be subsumed under the DN model, as the example discussed at the end of Sect. 2 demonstrates. In addition, none of those who insist on the importance of explanatory principles in life sciences (including cognitive science) have ever referred to the DN model. Because of that it is hardly reasonable to argue that the proponents of principle-based explanations in these fields are influenced by the DN model and/or not familiar with the real scientific practice.

(4) The case study of the basic level effects presented in Sect. 3 provides evidence which has been overlooked by the proponents of the mechanistic explanatory project: that some cognitive phenomena, like the basic level effects, do not seem susceptible to mechanistic explanations.

This paper does not aim to undervalue the mechanistic explanatory project. As Chater and Brown (2008) stated, a large class of cognitive phenomena could probably be explained by revealing the mechanisms that underlie/produce them. This paper merely draws attention to the fact that not all cognitive phenomena belong to that class. For these phenomena the principle-based explanations seem to be the only available choice.

References

Beatty, J. (1995). The evolutionary contingency thesis. In G. Wolters & J. Lennox (Eds.), *Concepts, theories, and rationality in the biological sciences* (pp. 45–81). Pittsburgh: University of Pittsburgh Press.

Bechtel, W. (2009). Constructing a philosophy of science of cognitive science. *Topics in Cognitive Science, 1*(3), 548–569.

Bechtel, W. (2010). How can philosophy be a true cognitive science discipline? *Topics in Cognitive Science, 2*(3), 357–366.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science, 78*(4), 533–577.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421–441.

Berlin, B. (1972). Speculations on the growth of ethnobotanical nomenclature. *Language in Society, 1*, 51–86.

Botstein, D. (1995). Structure and function of the gene. In M. Abeloff, J. Armitage, A. Lichter, & J. Niederhuber (Eds.), *Clinical oncology* (pp. 3–10). New York: Churchill Livingston.

Bouchard, F., & Rosenberg, A. (2004). Fitness, probability and the principles of natural selection. *The British Journal for the Philosophy of Science, 55*, 693–712.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65*, 14–21.

Chater, N., & Brown, G. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science, 32*, 36–67.

Crick, F. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology: The Biological Replication of Macromolecules, 12*, 152–153.

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*, 561–563.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). Cambridge, MA: The MIT Press.

Darwin, C. (1859). *The origins of species*. London: John Murray.

Fisher, R. (1930). *The general theory of natural selection*. Oxford: Clarendon Press.

Fodor, J. (2008). Against Darwinism. *Mind & Language, 23*(1), 1–24.

Fodor, J., & Piatelli-Palmarini, M. (2010). *What Darwin got wrong*. London: Profile Books.

Gibbs, R. (1994). *The poetics of mind: Figurative thought, language, and understanding*. New York: Cambridge University Press.

Katz, D. (1950). *Gestalt psychology*. New York: Ronald.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Mayr, E. (1985). How does biology differ from the physical sciences? In D. Depew & B. Weber (Eds.), *Evolution at a crossroads* (pp. 43–63). Cambridge, MA: The MIT Press.

Medawar, P. (1952). *An unsolved problem of biology*. London: H. K. Lewis.

Medin, D., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review, 111*, 960–983.

Morange, M. (2009). The central dogma of molecular biology. A retrospective after fifty years. *Resonance, 14*(3), 236–247.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: The MIT Press.

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*, 328–350.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale: Lawrence Erlbaum (Reprinted in E. Margolis, & S. Laurence (Eds.), *Concepts. Core readings* (pp. 189–206). Cambridge, MA: The MIT Press.).

Rosenberg, A. (2001). How is biological explanation possible. *The British Journal for the Philosophy of Science, 52*, 735–760.

Sarkar, S. (2005). *The molecular models of life: Philosophical papers on molecular biology*. Cambridge, MA: The MIT Press.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

Smith, E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Sober, E. (2008). Fodor's *bubble meise* against Darwinism. *Mind & Language, 23*(1), 42–49.

Stephen, D., & Van Orden, G. (2012). Searching for general principles in cognitive performance: Reply to commentators. *Topics in Cognitive Science, 4*, 94–102.

Van Orden, G., & Kloos, H. (2003). The module mistake. *Cortex, 39*, 164–166.

Wright, C., & Bechtel, W. (2007). Mechanisms and psychological explanation. In P. Thagard (Ed.), *Handbook of philosophy of science. Philosophy of psychology and cognitive science* (pp. 31–79). Amsterdam: Elsevier.

Computationalism, Connectionism, Dynamicism and Beyond: Looking for an Integrated Approach to Cognitive Science

Víctor M. Verdejo

Abstract Cognitive scientists are nowadays apparently required to choose between at least three different competing schools or general approaches: the computational, the connectionist and the dynamicist. More than three decades of unresolved paradigm fight encourage an alternative view: that each of these general approaches offer, not different explanations, but explanations of different aspects of cognitive phenomena. In this paper, I articulate this view by showing that each general approach can be taken to promote research primarily within a particular level of explanation. Failure to appreciate this fact has frequently led to largely incomplete accounts within each school. I argue that, if the articulation offered is sound, it supports the statement of an integrated programme for cognitive science where all the aforementioned general approaches have their place. Finally, I illustrate this analysis via a central theme for a clash of rival explanations in cognitive research, namely, systematicity.

Cognitive science is a discipline in continuous evolution where different and conflicting research strategies are permanently brought to the fore. As a consequence of discussion in the last 30 years or so, cognitive scientists are now apparently required to choose between at least three different schools or overall approaches: the computational, the connectionist, and the (embodied) dynamicist. By quite general assent, these approaches are understood as being rival views on cognition. It has seemed therefore fair to describe this situation in roughly Khunian terms (e.g. Schneider 1987; Chemero 2009): cognitive science cannot consistently be taken to include all these different approaches. On the contrary, each school stands for a

V.M. Verdejo (✉)
Universidade de Santiago de Compostela, Praza Mazarelos s/n,
15782 Santiago de Compostela, Spain
e-mail: vmverdejo@gmail.com

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_33,

different paradigm and the competing explanations they provide involve a typical scientific antagonism, one that results in a sort of radical paradigm fight. The conceptual and empirical discrepancies are so deep that, after this fight is resolved, at most one of these general approaches could turn out to be correct. The resulting view is somehow puzzling and disappointing. After many years of hard work in the cognitive sciences we might yet lack a clear answer to the question: what are the central theses that define cognitive science as a genuine discipline?

In this paper, I wish to make plausible the idea that this way of seeing cognitive research and the different approaches appeared in its wake is not only wrong but also itself highly pernicious for the correct assessment of its merits and achievements. Even though computationalism, connectionism and dynamicism are frequently seen as exclusive alternatives, much clarification can be gained by studying to which extent they articulate different aspects of one and the same scientific programme.[1]

This idea is certainly not new. Several authors have proposed analyses of cognitive science that, on the one hand, agree that computationalism, connectionism and dynamicism should be seen as three schools that may be united within a single framework (e.g. Cordeschi and Frixione 2007; Dawson 2013). On the other hand, other authors have actually developed specific aspects of this general idea in different fields, such as implementational connectionism (e.g. Marcus 2001), the theory of vision and visualizing (e.g. Pylyshyn 2003) or the representational approach to dynamicism (e.g. Grush 2003).

In this paper, however, I wish to explore the possibility of an integration of computationalism, connectionism and dynamicism in two steps: First, I will argue that in rough outline, the computational, the connectionist and the dynamicist models are plausibly considered as promoting lines of research focused and perhaps frequently 'misfocused' on one particular level of explanation. More precisely, I shall defend that these models correspond roughly to investigation developed primarily within one of Marr's three levels distinction (Marr 1982). Secondly, I will show that, if this is true, there is little reason to think that the aforementioned approaches to cognitive science are incompatible with each other. To the contrary, they are better seen as different lines of research belonging to a unified and integrated general inquiry into the nature of the mind.

Here is the route of the discussion to follow. In the first section, I outline the fundamental traits of Marrian explanation and present the view that computationalism, connectionism and dynamicism focus, or focus primarily, on research at one of these levels. In the second section, I show how concentration on one of Marr's levels has easily led in many cases to the problem of incomplete accounts in cognitive research. Section 3 is devoted to present what I, following Marr's own understanding of his

[1] Theoretical precision might require broadening the analysis to other cognitive schools as well. Extended, enacted, embedded, situated and distributed models might call for specific attention. However, simplicity and also the presumption that these other models might be reduced to or reasonably included into the ones here considered justify restricting our discussion to computationalism, connectionism and dynamicism.

tripartite distinction, present as integrated approaches in cognitive research, that is, approaches that take into consideration the complexity of cognitive phenomena at all levels of explanation. I finally illustrate, in the fourth section, the possibility of such an analysis with one central case in the history of cognitive studies, namely, the explanation of systematicity.

1 Different Approaches or Approaches at Different Levels?

Theorizing in cognitive science has probably so many perspectives as scholars working in the field. Nonetheless, it seems to be unobjectionable that, currently, there are at least three different and allegedly incompatible general strategies that may be held once engaged in the project of the cognitive study of the mind. On the one hand, there is the classical computational approach. This kind of approach capitalizes on the hypothesis that the mind is like a digital computer with discrete compositional symbols as mental representations (Newell and Simon 1972; Newell 1980). An alternative general approach emerges from consideration of connectionist networks. Connectionism challenges the classicist computational approach via the postulation of context-dependent and distributed mental representations in biologically plausible networks (McClelland and Rumelhart 1986; Rumelhart and McClelland 1986). A third and more recent alternative is found in so-called dynamicism or dynamical systems theory. This general approach takes the mathematical models of dynamics as the paradigm of cognitive research and thus provides explanations in terms of nonlinear differential equations (Horgan and Tienson 1994; van Gelder 1998).

I think it is plausible to consider the foregoing general approaches as focusing on and promoting research within one of Marr's levels (Marr 1982, chap. 1). For many years now, computational research has been used to the idea that there are at least three different levels of cognitive description or explanation. This is the seminal idea introduced by David Marr's influential computational account of vision. The three levels in question are, first, the one that specifies what the system does and why (Marr's level 1). Secondly, there is the level that states the algorithm and the representation (Marr's level 2). Finally, the lowest level takes into consideration the realization of the system in hardware structures (Marr's level 3). Remarkably, whereas much has been questioned in cognitive research, the explanatory importance of Marr's level-distinctions is, ultimately, pretty much untouched.[2]

[2] Marr was not alone in distinguishing levels of explanation in cognitive research. For instance, Newell (1986) or Pylyshyn (1984) provided a classification of cognitive explanations very similar to the one introduced by David Marr. These days, levels of explanation are, although largely unquestioned, also largely unattended to in the literature. This is arguably an unfortunate feature of contemporary cognitive theorising (see Verdejo and Quesada 2011 for an illustration of this point).

In this paper, I present an analysis of the aforementioned overall approaches in terms of Marrian levels so as to make plausible that, under certain standard readings, these approaches (1) centre research predominantly within one of these levels, taking the other levels to be irrelevant or else secondary; (2) as a consequence of (1), they often lead to incomplete and flawed accounts of cognition. Point (2) shall be the issue of the next section. In the remainder of this section, I will offer some considerations in favour of the plausibility of (1). Computational, connectionist and dynamical approaches can be analysed in terms of many different characteristic features. In what follows, I will try to show that at least some of these commonly attributed essential or distinctive features are explained because or appealing in the light of the kind of research to be expected at one of Marr's levels.

The introduction of (Marrian) levels of explanation for the analysis of computationalism, connectionism and dynamicism has been exploited by other authors as well (see e.g. Horgan and Tienson 1994; Cordeschi and Frixione 2007 and Dawson 2013). These analyses might differ, even greatly, with respect to the one presented here. It is important to note however that the existence of alternative analyses is no threat to the here presented line of argument. The argument only requires that, in the light of some essential or distinctive traits of each school, it is plausible to locate their contribution as belonging primarily to one of Marr's levels. It is only to be expected however that each of these schools (a) also exhibit commitments and substantial lines of research concerning other levels of explanation, and (b) may be judged as belonging to other levels in the light of different traits.

To begin with, then, let us discuss computationalism. Several ideas might be taken to be central to the models that characterize this school: serial processing, modularity, internal representations or an input-output analysis. Many, if not all, of these theoretical ingredients inevitably lead to the postulation of discrete symbolic structures on which different operations are defined. On the computational approach, cognitive states and processes are essentially defined in terms of the manipulation of these symbolic structures.

But note that discrete symbolic structures are precisely what one finds at high level characterizations of cognitive functions corresponding to Marr's level 1. This is especially clear if one considers that language and logical calculus (as opposed to e.g. perception and action) were the central concerns for the development of early cognitive research inspired by Turing-von Neumann architectures. Discrete symbolic structures, such as the ones found in the formulation of linguistic and logical problems (Newell and Simon 1972), are the hallmark of cognitive theorising at Marr's level 1. If we take cognitive science to consist of, or consist primarily of research at this level, it is only to be expected that the representational and the implementational level are taken to involve discrete representational and physical structures of the kind that result in a physical symbol system (Newell 1980). Whereas it is of course part of the computational view that there is an algorithmic and physical implementation of level-1 categories, it is nonetheless assumed that the structures identified at level 1 are entirely analogous to or provide the heuristics for the ones found at lower (algorithmic and physiological) levels.

Let us now move to connectionism. As is known, connectionist approaches appeared as a reaction to the computational dictum by emphasizing such things as the importance of fast, brain-like parallel processing, graceful degradation in the context of noisy inputs, the appeal to statistic and environmental data or the merits of associationistic mechanisms. The central tenet of connectionist models, however, has to do with the recourse to artificial neural networks or parallel distributed processing networks. Indeed, all of the aforementioned features are better seen as a consequence of the endorsement of this kind of networks as the core aspect of cognitive theorizing.

The adoption of connectionist networks is, patently, a fundamental assumption about the appropriate kind of representation, namely, distributed, multilayered and sub-symbolic representation (Smolensky 1988). But then, this is an alternative to the classicist computational model from the point of view of Marr's level 2. Connectionist models provide flexible context-dependent representations, which may be structurally very different from the categories identified at level 1 and which are arguably better suited for specific cognitive tasks such as fast pattern-recognition or unsupervised learning. On the other hand, if we put our bets on the algorithmic level of explanation as the genuine source for the understanding of cognitive phenomena then we might expect that the adoption of connectionism is or must be *definitive* regarding aspects certainly belonging to other levels, such as the endorsement of unstructured and purely context-sensitive functions (a claim at Marr's level 1) or biological plausibility (an implementational requirement at level 3).

The newest trend in cognitive science is dynamicism or dynamic systems theory. As with connectionism and computationalism, several aspects of this school might be emphasized: stability or self-organization of cognitive systems, agent-environment coupling or the quantification and nonlinearity of the variables relevant in the analysis. The fundamental trait of this kind of approach, however, is the statement of mathematical explanations of behaviour in terms of sets of differential equations that describe the evolution of a cognitive system in a state space or dynamic field.

It is natural to find this kind of approach appealing on the assumption that cognitive research has to be pursued primarily at Marr's implementation level 3. For one thing, according to such an assumption, cognitive systems are understood as being physical systems that continuously evolve over time. The assumption is precisely that cognitive systems are not different in nature (although they are certainly different in complexity) from a purely physical system –such as Watt centrifugal governor (van Gelder 1995). It is because a cognitive system is assumed to be a fundamentally physical system that the mathematics and dynamical language of physics is taken to be its appropriate explanatory tool.

For another thing, it is only in the light of such an assumption that we may find very natural indeed to endorse central ideas that accompany standard statements of dynamicist approaches (e.g. Calvo Garzón 2008; Chemero 2009). Two examples of these central ideas are, on the one hand, embodiment – i.e. the thesis that physical nature (as opposed to abstract functional analysis) is constitutive of cognitive systems–, and on the other hand, anti-representationalism – i.e. the view that representations at the algorithmic level have a residual explanatory role in cognitive research. If explanations at the implementational level are *the* genuine kind of explanations, then it follows that specifications of functions (at level 1) and of representations (at level 2) are dispensable or secondary. From this general point of view, dynamic models are naturally located within Marr's level 3.[3]

2 The Problem of Incomplete Accounts

Alas, and as I will be arguing in what follows, the fact that computationalism, connectionism and dynamicism focus on a particular Marrian level turns, more often than not, on a 'misfocus' as regards *bona fide* cognitive research. More precisely, it is easy to see many aspects of the dialectics confronting these schools as arising from the incompleteness of their accounts regarding the rest of levels of explanation.

Thus, for instance, computational models crystallized in Jerry Fodor's (1975, 2008) seminal Language of Thought Hypothesis (LOT henceforth). Frequent interpretations of this approach however, overemphasize the characterization of (specifically linguistic) cognitive processes from a high-level point of view, indeed, from a level at which common sense psychology could be resolutely vindicated (Fodor 1987). The idea was simple as it was controversial: if a computer parses sentences via a programming language, such as List Processing programming language (LISP), then a mind must also comprehend sentences via some still to be specified programming mental language. Standard developments of this idea easily lead to the impression that consideration of research at the algorithmic and implementational levels is subsidiary. The rule of thumb appears to be to extrapolate psychologically real representational categories and physical symbols from rough characterizations of grammatical parse-dependent functions. From this perspective, and although a LOT strategy arguably amounts to a general cognitive model applicable to a variety of domains (Verdejo 2012), it may seem that the classicist computational approach to cognition is confined to abstract linguistic or logical operations, and is therefore empirically unwarranted and insensitive to the pragmatic, contextual and physical concomitants of cognition (Dreyfus 1992; Hurley 2001; Chemero 2009). In short, much of the dialectical opposition to computational approaches can be seen as

[3] Admittedly, some authors have argued that dynamicism can be seen as contributions to Marr's level 1 (Cordeschi and Frixione 2007) or the algorithmic level 2 (Horgan and Tienson 1994). These approaches take as a defining feature of dynamical systems the statement of mathematical models from an abstract point of view. However, this interpretation of dynamicism arguably leaves unexplained the fundamentally embodied and anti-representational character of recent developments. This is not the place to develop this claim further. For present purposes, it suffices that the location of this school at Marr's level 3 is plausible in the light of the dynamicist traits just mentioned in the main text.

stemming from the mistake of considering, however tacitly, that serious research at the algorithmic and physical level is secondary or dispensable.

Now, as noted, it was a central part of connectionism to resist classicist computational models by focusing on Marr's level 2 and offering non-purely-linguistic or sub-symbolic alternatives (Smolensky 1988). However, the 'misfocus' of connectionism on the algorithmic level generally prevented the approach to properly capture the importance of level-1 characterizations. This was used by computationalists to mount a long-lived challenge to connectionism regarding systematicity (Fodor and Pylyshyn 1988): either connectionism does not explain the systematicity functions or else it explains them only as an implementation of computational models. The source of this and similar challenges is the absence of a clear statement of the function, if any, connectionism was showing how to compute.

On the other hand, a paradigmatic criticism of connectionism stems from its inability to account satisfactorily for the relation between connectionist and neuronal, psychologically and physiologically real networks (e.g. McLaughlin and Warfield 1994). In this respect, predominant connectionist advertising has failed to see that, qua contribution to the algorithmic level of explanation, the neuron-like, implementational virtues were often *not* a reasonable requirement for connectionist networks.

Finally, dynamic mathematical models, originally presented as belonging to the same family as connectionist models (van Gelder 1995), have engendered original explanations of traditional cognitive problems in perception, developmental psychology, sensorimotor tasks and much else. Unfortunately, focus or 'misfocus' on issues at Marr's level 3 has also brought dynamicism into a picture where, at the algorithmic level, anti-representationalism per se is an (empirically doubtful) scientific desideratum (Grush 2003). As a consequence of this, and as in some ramifications of the connectionist school, the absence of clear level-1 characterizations inevitably leads to neglect many high-level categories and to encourage some form of controversial behaviourism or eliminativism (e.g. Ramsey 2007).

3 An Integrated Approach to Cognitive Science

It seems to me that if the foregoing characterization is roughly correct, it automatically suggests an alternative conception of cognitive science where (a) the problem of incomplete accounts of cognitive phenomena does not arise in the first place and (b) computationalism, connectionism and dynamicism are prima facie compatible with each other after all. Partial views in cognitive research that focus or 'misfocus' on a particular level of explanation can be plausibly amended by views in which all levels of explanation are integrated. In fact, the possibility of an integrated account of cognition was arguably what Marr had in mind when he distinguished between levels of explanation in the first place.

Note that, according to Marr, his distinction responds to the real complexity of cognitive phenomena. Thus, the levels must be carefully attended to because "explication of each level involves issues that are rather independent of the other two" (Marr 1982, p. 25). More importantly in the context of this paper, all three levels are levels "at which an information processing device must be understood before one can be said to have understood it completely" (Marr 1982, p. 24). Marr's fundamental contention was therefore not that one particular level is methodologically or explanatorily prior to the others but that full explanation of cognitive phenomena must involve correct explanations at all three levels. Unsurprisingly then, Marr thought that even the highest level 1, the level of what he called the computational theory, is a level at which we "should look out for the physical constraints [at level 3] that allow the process to do what it does" (Marr 1982, p. 103).

What these considerations suggest is that the best insights of computationalism (regarding level 1), connectionism (at level 2) and dynamicism (at level 3) may be integrated and thus make justice to the extraordinary complexity of cognitive phenomena via an extraordinarily rich repertoire of explanatory models. This is, to be sure, easier to say than to carry out. For this reason, I will introduce in the final section of this paper an illustration of how this guiding idea might go as regards a long-standing dialectic in cognitive research, that is, the systematicity debate.

4 The Case of Systematicity

Let us now consider, as an illustration of the view here presented, the case of an integrated account of systematicity. The illustration will be divided into two stages. On the one hand, I will introduce the fundamental explanatory schema of systematicity phenomena generally. At a second stage, I will present a series of considerations to the effect that an integration of cognitive models is clearly possible and certainly desirable in such a case.[4]

To a first approximation, systematicity is that property in virtue of which if a subject S has a given capacity C, then as a matter of fact, S has another, structurally related, capacity C'. The legendary example is that if S has the capacity to produce/understand the sentence "Mary loves John" then S, as a matter of fact, has also the capacity to produce/understand the sentence "John loves Mary". Standard doctrine takes it that the correct explanation of systematicity involves compositionality (Fodor 1987, 2008; Fodor and Pylyshyn 1988). In other words, in order to provide a correct explanation of systematic states and processes one has to state, satisfactorily, a constituency relation between a corpus of primitive elements and the corresponding complexes made out from these primitives. Now, a first step in showing the possibility of an integrated approach in this case is

[4] Some authors will be willing to cast doubts on systematicity itself. In what follows, however, I will analyse the explanations that each school provides of the phenomenon and therefore I will assume that systematicity is an (empirically confirmed) explanatory target for each of the schools under analysis.

to note that constituency relations, of the sort that explain systematicity, can be consistently stated among different kinds of elements, depending on the level at which our explanation is operating.

Apparently, however, the explanations of systematicity phenomena that computationalism, connectionism and dynamicism have favoured exhibit radically opposing views. Classical computationalism has taken systematicity as perhaps the strongest evidence for the existence of a LOT, a linguistic representational system of discrete symbolic elements apt for combination and recombination (Fodor 1987; Fodor and Pylyshyn 1988). On the other hand, connectionist theorizing found that alternative, non-linguistic kinds of representational schemes –such as tensor-product networks (Smolensky 1990; Smolensky et al. 1992) or Gödel numbering (van Gelder 1990) – could provide all the sensitivity to structure that systematicity demands. Embodied dynamicism, finally, went on to postulate categories within dynamic theory in order to articulate a suitable interpretation of systematicity. The key idea is that systematicity is a mathematical relation between points or basins of attraction in state space (Horgan and Tienson 1994; Calvo Garzón 2004).

Considered in isolation, the resulting accounts within each school constitute a vivid example of the problems of incomplete accounts. First, misfocus on the linguistic structures that allowed the explanation of systematicity at Marr's level 1, made computational approaches end up with the implausible commitment that algorithmic and neurophysiological counterparts of those structures should be actually 'linguistic' in a very unpalatable sense. This was expressed sometimes as the requirement that systematicity-explaining constituency relations involved a sort of concatenative or spatio-temporal relation among constituents (e.g. Fodor and McLaughlin 1990), as if LOT was in an implausible literal sense necessarily a written or spoken language at the algorithmic and physiological levels.

To great avail, connectionism introduced the alternative sub-symbolic versions of representations for the explanations of systematicity. It was proved mathematically that classical explanations were equivalent to subsymbolic kinds of representations such as tensor-product networks (Smolensky et al. 1992). However, proponents of connectionist networks usually fell prey to a wrong assessment of their real findings: they were not offering alternative general models of cognition at all levels but, quite differently, alternative models at the algorithmic Marrian level 2 for already specified systematicity functions.

Finally, the anti-representational and embodied versions of dynamicism led to the idea that systematicity should be treated as a physical process accounted for by appeal to the mathematical tools and language of differential equations (Calvo Garzón 2004). The dynamicist proposal however introduced considerable confusion as to what kind of systematicity these models are actually explanations of –what the systematicity function they are modelling is – and left mysterious the real and obvious representational demands at level 2 of systematicity phenomena at large.

Careful scrutiny however can show that all these competing explanations of systematicity are only superficially incompatible. Take the 'Mary-loves-John' case. At the level of the function that the system computes (Marr's level 1), the problem and the explanatory categories are obviously linguistic in nature. What we identify at this level is a pair of systematically related capacities, namely, the capacity to understand/produce "Mary loves John" and the capacity to produce/understand "John loves Mary". At this level, constituency relations of the sort that explain systematicity are stated between words or concepts and sentences or propositional thoughts.

At the level of the algorithm, however, we do not need to appeal to the very same categories that explained the phenomenon at the level of the function. It suffices that the representations postulated at this level are actually representations relevant for the explanation of how the system performs the function identified at level 1. A number of algorithmic possibilities would be available. These include programming languages and vector binding. It is not necessary that the representations involved mimic the very same linguistic structure appealed to in the explanation of the phenomenon at the highest level.

Finally, the continuous physical evolution of the system at the lowest implementational level may leave room for a dynamical insight in terms of nonlinear differential equations that define basins of attraction or a series of points in state space. But the consideration of this possibility does not exclude the existence of linguistic explanatory categories at level 1 or representational categories at level 2. Indeed, these accounts can be seen as showing central dynamical aspects of the way in which level-1 and level-2 categories are physically implemented.

Now we are in a position to see clearly what is gained by adopting an integrated approach in the case of systematicity. In an integrated approach, the constituency relations relevant for the explanation of systematicity are highlighted as involving different sorts of facts at different levels. The full complexity of cognitively relevant constituency relations is thus uncovered and carefully analysed. On an integrated account, constituency relations may be consistently stated between words or lexical categories (at level 1), activation vectors (at level 2) and points or basins of attraction in a dynamical space (at level 3). In the terms I am proposing, this means that an integration of the basic explanatory models under study –the computational, the connectionist and the dynamicist– is clearly possible.

In addition, one such integration is arguably highly desirable. The combination of rigorous accounts at different levels may serve to rectify, contrast and constrain the validity of assumptions at every level. For instance, computational models are presumably right to insist that it is compositional symbols, that is, semantically-cum-syntactically characterized representations that explain systematicity. What advocates of computationalism usually neglect, however, is that since what we are concerned with is cognitive and, to that extend, psychologically real compositionality, a correct statement of the corresponding representations is certainly more involved than, say, high-school grammatical analysis. Algorithmic and physical levels have to be carefully attended to.

The available algorithms, in turn, must take into consideration the full range of phenomena that count as systematicity phenomena (Cummins et al. 2001), and hence include, for instance, the systematicity found in vision and visual abilities. This means that the systematicity algorithm and representation is very likely of a connectionist kind (as opposed to a programming language kind) and will never be 'linguistic' in a very literal sense (Verdejo 2012).

Finally, the physical constraints of the organism, plausibly modelled in dynamic terms, have a substantial role, on the one hand, in the specification of physically plausible mechanisms. Functions and algorithms that the organism computes will be translated into dynamical systems with their own restrictions regarding relevant physical variables and inter-dependence among those variables. On the other hand, the anti-representational leanings of dynamical systems (Calvo Garzon 2008) may prevent the postulation of unnecessary representational categories and hence contribute to a more demanding specification of their explanatory role.

In sum, when systematicity is considered as an 'integrated' cognitive phenomenon, explanations in terms of compositionality and constituency relations result in a substantial, and substantially complex, inter-level account. This is precisely what we get if we buy in for an approach to systematicity of the sort here defended.

5 Conclusion

In this paper, I have undertaken the task of showing, in broad and brief outline, that there is a promising alternative to the opposing terms in which different approaches to cognitive science are usually presented. The alternative suggests that many central aspects of the unresolved confrontation between computationalists, connectionists and dynamicists are better seen as a failure to appreciate the integrative relations between researches at different levels of explanation. From these considerations, the sketch of an integrated approach emerges, one in which empirically grounded explanation at different Marrian levels results in a powerful and rich framework for cognitive science.

Acknowledgement I am grateful to the audience of the EPSA11 conference in Athens and to an anonymous referee for their valuable comments and suggestions on earlier drafts of this paper. This research work has received financial support from the Ministerio de Economía y Competitividad (Spanish government), via the research project FFI2009-08828/FISO, and from the Generalitat de Catalunya (Catalan government), through the consolidated research group GRECC (SGR2009-1528).

References

Calvo Garzón, F. (2004). Context-free versus context-dependent constituency relations: A false dichotomy. In S. Levy & R. Gayler (Eds.), *Proceedings of the American Association for Artificial Intelligence* (pp. 12–16). Menlo Park: AAAI Press.

Calvo Garzón, F. (2008). Towards a general theory of antirepresentationalism. *British Journal for the Philosophy of Science, 59*, 259–292.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Cordeschi, R., & Frixione, M. (2007). Computationalism under attack. In M. Marraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the mind* (pp. 37–49). Dordrecht: Springer.

Cummins, R., Blackmon, J., Byrd, D., Poirier, P., Roth, M., & Schwarz, G. (2001). Systematicity and the cognition of structured domains. *Journal of Philosophy, 98*, 167–185.

Dawson, M. R. W. (2013). *Mind, body, world: Foundations of cognitive science*. Edmonton: Athabasca University Press.

Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge, MA: MIT Press.

Fodor, J. A. (1975). *The language of thought*. New York: Thomas Y. Crowell.

Fodor, J. A. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. A. (2008). *LOT2: The language of thought revisited*. Oxford: Oxford University Press.

Fodor, J. A., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition, 35*, 183–205.

Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

Grush, R. (2003). In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Phylosophy, 18*, 53–93.

Horgan, T., & Tienson, J. (1994). A nonclassical framework for cognitive science. *Synthese, 101*, 305–345.

Hurley, S. (2001). Perception and action: Alternative views. *Synthese, 129*, 3–40.

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel distributed processing* (Vol. 2). Cambridge, MA: MIT Press.

McLaughlin, B., & Warfield, T. (1994). The allure of connectionism re-examined. *Synthese, 101*, 365–400.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135–183.

Newell, A. (1986). The symbol level and the knowledge level. In Z. Pylyshyn & W. Demopoulos (Eds.), *Meaning and cognitive structure* (pp. 31–39). Norwood: Ablex.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs: Prentice Hall.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2003). *Seeing and visualizing*. Cambridge, MA: MIT Press.

Ramsey, W. (2007). *Representation reconsidered*. New York: Cambridge University Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Schneider, W. (1987). Connectionism: Is it a paradigm shift for psychology? *Behaviour Research Methods, Instruments, and Computers, 19*, 73–83.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioural and Brain Sciences, 11*, 1–74.

Smolensky, P. (1990). Tensor product variable binding and the representation of structure in connectionist systems. *Artificial Intelligence, 46*, 159–216.

Smolensky, P., Legendre, G., & Miyata, Y. (1992). Principles for an integrated connectionist/symbolic theory of higher cognition. Institute of Cognitive Science, University of Colorado, Technical Report 92-08.

Van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science, 14*, 355–384.

Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy, 92*, 345–381.

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioural and Brain Sciences, 21*, 615–628.

Verdejo, V. M. (2012). Meeting the systematicity challenge challenge: A nonlinguistic argument for a language of thought. *Journal of Philosophical Research, 37*, 155–183.

Verdejo, V. M., & Quesada, D. (2011). Levels of explanation vindicated. *Review of Philosophy and Psychology, 2*, 77–88.

Qualia Change and Colour Science

Lieven Decock and Igor Douven

Scepticism is an offshoot of science. The basis for scepticism is the awareness of illusion, the discovery that we must not always believe our eyes.

W. V. O. Quine (1975, p. 67)

Abstract Many contemporary qualia inversion arguments are inspired by findings in colour science, most notably, the Hering-Jameson-Hurvich opponent processes theory. This is somewhat ironic, given that other findings in colour science – particularly findings indicating that phenomenal colour space is asymmetrical – appear to exclude qualia inversion scenarios. In previous work, we proposed an alternative qualia change scenario – called "qualia compression" – that is impervious to the asymmetry objection. The present paper argues that qualia compression is more than merely another thought experiment. We do this by connecting it to recent developments in colour science. Specifically, we point at experiments on gamut expansion and compression by Brown and MacLeod, Li and Gilchrist, and Whittle.

L. Decock (✉)
Faculty of Philosophy, VU University Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands
e-mail: l.b.decock@vu.nl

I. Douven
Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands
e-mail: j.douven@rug.nl

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_34,
© Springer International Publishing Switzerland 2013

1 Qualia Inversion and Colour Science

John Locke may have been the first to put forward the idea that different people may experience different colour qualia (in *An Essay Concerning Human Understanding*, II, xxxi, 15). Specifically, he wondered whether one person might experience marigolds by the “idea” yellow and violets by the “idea” blue, while another experiences marigolds by the “idea” blue and violets by the “idea” yellow. Because he believed that this thought experiment was “of little use either for the improvement of our knowledge or conveniency of life” (*ibid.*), Locke quickly dismissed it – too quickly, according to many later authors, who have invoked it, or variants of it, in epistemological discussions on scepticism and ontological debates concerning the nature of perceptual experience.[1] In fact, nowadays we mostly encounter versions of the thought experiment that are more elaborate and also more sophisticated than Locke’s, versions that postulate much more radical and systematic changes in a person’s colour experiences than Locke’s did, and often also appeal to experimental results in colour science.

The name now commonly in use for this family of thought experiments – “spectral inversion arguments” – is indicative of the type of change that is typically considered, to wit, the inversion of the experienced qualia of the colours of the spectrum. One precise formal model of spectral inversion can be formulated by reference to Newton’s theory of the diffraction of light. If one considers a rainbow, or the light diffracted by a prism, it is imaginable that one person experiences light beams with wavelengths of 400 nm as violet and beams with wavelengths of 700 nm as red, while someone else experiences them inversely, in the sense that, for any wavelength $\nu \in [400\text{ nm}, 700\text{ nm}]$, the quale that in one person is elicited by a light beam of wavelength ν is identical to the quale that is elicited in the other person by the wavelength $(1{,}100 - \nu)$ nm. This inversion scenario, however, fits badly with several findings in colour science. It has long been known that colour hues are better represented on a colour circle (or square, or hexagon) than on a line fragment. This is because there is a continuum of purple shades between violet and red that cannot be found in the diffraction spectrum. Moreover, the colour shades in the colour spectrum are fully saturated, while many colour shades – such as brown and olive – are not. Thus, spectral inversion of the above type can yield a systematic qualia change for at most a limited class of colour shades.

[1] For an overview of these discussions, and for the role qualia change scenarios play in them, see Byrne (2012). Byrne also contains valuable discussion of various responses to scepticism engendered by qualia change scenarios. The qualia change scenarios considered in this paper are meant to address more specific criticisms of qualia change. In particular, we will not be concerned with criticisms that call into doubt the existence of qualia or the possibility of shared qualia, or that argue for a general operationalism, which would render the notion of undetectable qualia change void.

More refined versions of colour qualia inversion arguments tend to rely on the systematic organization of colours in a phenomenal colour space. Colour space is generally assumed to be a three-dimensional Euclidean space, with one dimension representing hue, one representing saturation or chromaticity, and one representing lightness. The hue dimension can be thought of as a circle with the spectral colours red, orange, yellow, green, blue, violet (neighbouring red again) lying on it, where one colour gradually merges into the next. Both saturation and lightness are linear dimensions with a minimum and a maximum. Saturation indicates the intensity or fullness of a colour and lightness ranges from black to white, going through all shades of grey. Phenomenal colour space encodes several important structural relations between colour qualia, most notably, similarity, mixing, and opponency: qualia at close distance in phenomenal colour space are experienced as highly similar, and below a certain threshold, they may even be indistinguishable; mixing two colours with a particular hue will yield a resulting hue in between the two mixed colours (which, note, requires that phenomenal colour space is convex); and opponent colours, such as red and green, are on opposite sides of colour space. Many of the more recent qualia inversion scenarios postulate an inversion of opponent colours.

Colour opponency was first described in Goethe's ground-breaking work on the phenomenology of colour vision and was later studied more thoroughly in the work of the psychologist Hering. Hering claimed that red, blue, green, and yellow are the four basic colours, produced by two similar antagonistic processes. At the end of the nineteenth century, Hering's ideas were incorporated in Höfler and Ebbinghaus's representations of phenomenal colour space as a double pyramid with opponent colours lying on the diagonals of a square. Jameson and Hurvich (1957) expanded on the theory by proposing three channels through which information about differences in electromagnetic activation of the three types of cones in the retina – the **L**-, **M**-, and **S**-cones – is processed through the optical nerve, thus obtaining a quantitative opponent process theory. There has been continued debate over the precise form of the opponent processes, but there is considerable consensus over the following model (Hardin 1988, p. 35):

- $(\mathbf{L}+\mathbf{M})$ is the achromatic signal:
  - $(\mathbf{L}+\mathbf{M}) > 0$ codes whiteness;
  - $(\mathbf{L}+\mathbf{M}) < 0$ codes blackness;
  - $(\mathbf{L}+\mathbf{M}) = 0$ codes for brain grey.
- $(\mathbf{L}-\mathbf{M})$ is the red-green channel:
  - $(\mathbf{L}-\mathbf{M}) > 0$ codes for redness;
  - $(\mathbf{L}-\mathbf{M}) < 0$ codes for greenness.
- $(\mathbf{L}+\mathbf{M})-\mathbf{S}$ is the yellow-blue channel:
  - $(\mathbf{L}+\mathbf{M})-\mathbf{S} > 0$ codes yellowness;
  - $(\mathbf{L}+\mathbf{M})-\mathbf{S} < 0$ codes blueness.

In this model, the achromatic signal is based on the degree of activation of the **L**- and **M**-cones, while chromatic vision is based on differences in activation of all three types of cones. Although some mathematical particulars must be filled in – for instance, the sign is conventional and weighting coefficients may be needed – the opponent processes model almost evokes the qualia inversion scenarios. For note that by changing the sign of $\mathbf{(L-M)}$ or $\mathbf{(L+M)-S}$, we obtain a red-green or a blue-yellow inversion, respectively.[2] An extra inversion is obtained by interchanging the $\mathbf{(L-M)}$ and $\mathbf{(L+M)-S}$ coordinates, thus obtaining a red-yellow and green-blue inversion.

The possibility of qualia inversion hinges on the symmetry of phenomenal colour space. If the $\mathbf{(L-M)}$ channel and the $\mathbf{(L+M)-S}$ channel are represented by the main axes of a colour circle in a colour spindle or by the diagonals of a colour square in a double pyramid, the inversions map every point in phenomenal colour space onto another point of the space, and the unique hues are mapped onto unique hues, thus producing colour inversions. However, it is generally believed that phenomenal colour space is *not* symmetrical. For example, unique yellow is believed to be less saturated than unique red.[3] Indeed, ever since Munsell, colour scientists trying to faithfully represent phenomenal colour space have come up with asymmetrical colour appearance models. Hence, Hardin's (1988, pp. 134–142) conclusion that qualia inversion is impossible, a conclusion that is now widely accepted (cf. Byrne 2012).

2 Qualia Compression

The asymmetry objection to qualia inversion scenarios can in fact be stated without reference to the Hering-Jameson-Hurvich opponent processes theory. The detailed statement of the objection, which is found in Decock and Douven (2013), makes use of the fact that all similarity judgments can be expressed by means of a quaternary similarity relation Sim(w,x,y,z), which is to be interpreted as "w is more similar to x than y to z." (After all, ternary similarity judgments can be formulated as "x is more similar to y than x is to z," and binary similarity judgments can be expressed as "x is more similar to y than s is to $s*$," for two reference samples s and $s*$.) In phenomenal colour space, perceived similarities are represented as distances between points representing the qualia elicited by the stimuli. In particular, where Q_x is the quale elicited by stimulus x, the sentence "a is more similar to b than c is to d"

[2] Some philosophers have also considered inversions of the achromatic black-white channel; see Myin (2001) and Broackes (2007).

[3] Differently stated, the maximal positive value of $\mathbf{(L-M)}$ is greater than the maximal positive value of $\mathbf{(L+M)-S}$.

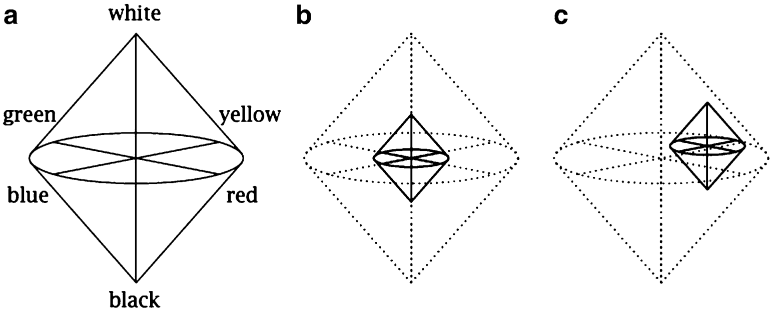


Fig. 1 The colour spindle (**a**); compression by rescaling the colour spindle (**b**); compression by rescaling and translating the colour spindle (**c**)

is true if and only if the distance in phenomenal colour space between Q_a and Q_b is smaller than the distance between Q_c and Q_d. On the assumption that phenomenal colour space is an asymmetrical three-dimensional Euclidean space, it is readily shown that no nontrivial automorphisms within phenomenal colour are possible that leave all similarity judgments intact. For example, since saturated yellow is closer to the centre of phenomenal colour space – known as "brain grey" – than saturated red is, a red-yellow inversion would alter the truth value of the similarity judgment: "This saturated yellow sample is more similar to this brain grey sample than to this saturated red sample."

In Decock and Douven (2013), we also showed that this formulation of the objection to qualia inversion leaves room for an alternative reshuffling of phenomenal colour space. Instead of considering an automorphism on this space, we considered an isomorphism between phenomenal colour space and one of its inner regions that preserves the relevant structure on which the similarity judgments are based.

To illustrate the basic idea, assume that phenomenal colour space has the form of a spindle, as depicted in Fig. 1a. Then a one-to-one correspondence between phenomenal colour space and one of its inner regions can be constructed by mapping every point onto a point that lies on a fixed ratio between it and the centre of the spindle. One may thus think of a compression as yielding an isomorphic miniature spindle that lies inside phenomenal colour space. In this case the miniature spindle has the same centre as phenomenal colour space (as in Fig. 1b). For every ratio between 0 and 1, we obtain a compressed spindle. In a next step, we can also consider translations of the spindle thus obtained (see Fig. 1c). For example, we can consider a compression which maps phenomenal colour space onto a miniature spindle lying in the pink area, with black mapped onto the darkest pink and white onto the lightest pink. Obviously, this compression procedure does not depend on the precise form of phenomenal colour space. In particular, asymmetries of that space pose no obstacle to compressions.

In Decock and Douven (2013), we made this idea more precise by formally defining a qualia compression in a given three-dimensional Euclidean space. We showed that any qualia compression can be expressed mathematically as a combination of a rescaling of the coordinate axes and a translation.[4] Specifically, we defined a function f that, for any point $a = \langle x,y,z \rangle$ in phenomenal colour space, yields the value:

$$f(a) = \langle rx, ry, rz \rangle + \langle x_o, y_o, z_o \rangle,$$

where $r (0 < r < 1)$ is a scaling factor expressing the degree of compression; x_o, y_o, and z_o are the coordinates of point to which the center of phenomenal colour space is translated; and the values of r and o are to be so chosen that the range of f lies entirely within phenomenal colour space. Letting d designate the metric on phenomenal colour space, we showed that the distance $d(f(a),f(b))$ equals $r \times d(a,b)$, that is, that distances in the compressed space are shrunk by a factor. On this basis, and using the characterization of the quaternary similarity relation stated above, it was further shown that for any colour stimuli a, b, c, and d, it holds that

$$\mathrm{Sim}\left(f(a), f(b), f(c), f(d)\right) \iff \mathrm{Sim}\left(a, b, c, d\right),$$

and thus that all similarity judgments are left invariant under any given qualia compression. We concluded from this that qualia compressions yield behaviourally undetectable changes of qualia and that, as a result, they can serve the same sceptical purposes as the qualia inversion scenarios were meant to serve. After all, an individual with compressed colour qualia will have a limited range of colour experiences – limited relative to a person with normal colour qualia – without him or her, or anyone else for that matter, ever being in the position to find out about this limitation, which is precisely what the sceptic wants to maintain.[5]

[4] An anonymous referee brought to our attention that, alongside with rescalings and translations, we could have considered rotations. While this is correct, it is to be noted that, as will be seen, the scientific literature provides evidence for the actual occurrence of rescalings and translations of phenomenal colour space, but not, to the best of our knowledge, for the occurrence of rotations of that space.

[5] Hardin (1988) discusses a number of asymmetries of phenomenal colour space in the context of arguing against qualia inversion arguments that might also seem problematic for our qualia compression argument. In particular, he points out that for colour qualia inversion to work, fully saturated hues must be mapped onto fully saturated hues, mapped shades must belong to the same Basic Colour Terms, and warm colour shades must be mapped onto warm colour shades. In Decock and Douven (2013), we showed that these possible objections are forceless if we think – as we plausibly should – of saturatedness, basicality, etc., as relational rather than intrinsic properties of qualia.

3 Qualia Compression and Colour Science

Even though the latest empirical results seem to tell against qualia inversion arguments, some of these arguments are, as we said, clearly rooted in empirical science. From the presentation in Decock and Douven (2013), one could get the impression that qualia compression is only a theoretical possibility, but in fact it has a basis in colour science. For in recent years several empirical findings have shown the possibility of so-called gamut compression or gamut expansion. In fact, both compressions and expansions of phenomenal colour experience can be found in the three dimensions of phenomenal colour space: along the saturation axis, the lightness axis, and in the hue dimension. These experiments, which we will now briefly discuss, show that distances in phenomenal colour space are not rigidly tied to similarities between physical stimuli; in some viewing conditions, the experienced similarities may appear smaller (gamut compression) or larger (gamut expansion) than the normally perceived similarity.

3.1 *The Saturation Axis: Brown and MacLeod*

Brown and MacLeod (1997) present surprising results concerning a phenomenon they coined "gamut expansion." Before their study, it was known that surround colours can shift the phenomenal experience of target areas. Typical demonstrations of this phenomenon rely on simultaneous colour contrast. For instance, it was known that when two identical grey discs are placed against different backgrounds, they no longer appear identical but seem tinged with a colour roughly complementary to that of their surrounds. It was generally believed that this phenomenon could be modelled by translations in colour space, the thought being that adding surround colours shifts the position of the "white point" in colour space. However, Brown and MacLeod discovered that some surround effects cannot be explained in terms of translations in colour space, but are more plausibly due to (local) expansions and/or compressions of colour space.

In Brown and MacLeod's experiment, observers were shown six rectangles that were predominantly grey but tinged with yellow, white, red, green, black, and blue, respectively. The rectangles were presented against two different backgrounds. The first background had a multi-coloured mosaic pattern, which could be either a chessboard pattern or (what Ekroll et al. 2011 called) a Seurat pattern (with reference to the celebrated French pointillist), consisting of small overlapping discs; the squares or discs in this surround were randomly coloured in shades with high colour contrast. The second background was uniform and middle grey. The two backgrounds were so chosen that they had the same space-averaged colour in order to exclude a translation of the white point. In the experiment, the colour differences among the six rectangles were experienced quite vividly against the uniformly grey background, but they were experienced as similar greyish shades in the multicoloured surround. The phenomenal effect could be described as an expansion[6] of colour space in the environment of the background colour. In particular, the slightly coloured shades were experienced as more saturated. In their paper, Brown and MacLeod submit that this could be a common but often overlooked phenomenon, attributable to a mechanism that tends to preserve a relatively large gamut of perceived colours in low-contrast viewing conditions (such as in a fog or a haze), and which achieves this by expanding experienced colour differences so that colours come to appear more saturated.

3.2 The Lightness Axis: Gilchrist

In a sense, the rescaling of the lightness axis is quite common, given that scaling the lightness axis is context-dependent. Because the cones in the retina can only detect relative differences in activation, the phenomenal experience of white and black must be "anchored" (Gilchrist 2006, Chap. 9); that is, for any particular visual scene, the lightness axis must be determined anew. This anchoring problem is not trivial. It is well known that luminance values in the proximal stimulus – or the corresponding degrees of activation of the cones in the retina – do not determine phenomenal lightness experiences. This luminance is largely dependent on the background illumination, whereas the perceived lightness is largely independent of the background illumination, as is for instance clear from the fact that the page of a book appears white both in a badly lit room and in bright sunlight. Moreover, relative illumination provides no better immediate clue to the perceived lightness. Knowing that the luminance of a target surface is five times the luminance of an adjacent surface tells us little about their perceived lightness. If the adjacent surface is perceived as black, the target surface will appear grey; if the adjacent surface appears grey, the target surface will appear white; and if the adjacent surface appears white, the target surface will appear self-luminous. In order to determine the perceived lightness of a target object, we will have to consider the relative lightness of all the objects and light sources in the environment. Hence, the determination of the lightness of an object is context-dependent and the lightness axis scaling is relative to the overall composition of the visual scene.

There is a fairly reliable rule – the so-called highest luminance rule – to anchor the lightness axis for any given visual field. According to this rule, the surface with the highest luminance is experienced as pure white. The luminance ratio between white and black equals . Object surfaces with a luminance below roughly 3 % of the white surface are experienced as pure black.

[6] The effect could also be due to a gamut compression in the multicoloured surround. However, Ekroll et al. (2011) provide evidence that this is less plausible than the interpretation mentioned in the text.

However, also this relativized lightness scale can be compressed in a "gamut compression" (Gilchrist 2006, pp. 236–239). In addition to the highest luminance rule, another principle is at work in lightness experience, namely, the largest area rule. In normal cases where the largest area in the viewing field is lighter than a smaller dark area, the two principles are not in competition. However, if the dark area is larger than the light area or areas, the largest area rule implies that an increase in area is perceived as an increase in luminance (within certain limits) and that the largest area has a tendency to appear white. As a result, if the dark area is the largest, we may witness gamut compression, that is, a compression of the perceived darkness.

A nice illustration of this type of phenomenon was found in an experiment by Li and Gilchrist (1999, Experiment 1). In this experiment, participants were asked to place their heads into a hemisphere the inside of which was painted partly black and partly middle grey. For one group of participants, the hemisphere was evenly split between a black area and a middle grey area; for a second group, the hemisphere was divided into a large black and a small middle grey area. In both conditions, participants perceived the middle grey area as white, in accordance with the highest luminance rule. The black area, on the other hand, was perceived differently in the two conditions: whereas in the evenly split condition, it was perceived as middle grey, in the unevenly split condition it was perceived as very light grey, in accordance with the largest area rule. Thus, an actual compression – in this case of the relative perceived lightness difference – was achieved by enlarging the black area and shrinking the middle grey area.

3.3 Hue Circle: Whittle

A striking example of hue expansion is found in Whittle (2003), whose central claim is that contrast effects have often been underrated in colour science. To buttress this claim, Whittle conducted an experiment in which a row of reddish shades and a row of bluish shades were experienced as matched and ranging over the full gamut of hues, thus yielding a dramatic hue expansion. Whittle arrived at this result by choosing two sets of colours in the constant luminance diagram (i.e., the logarithmic version of the MacLeod-Boynton diagram) in which all the colour shades of constant luminance are represented. He chose two reference colour shades, one in the red and one in the blue area. Around each of these reference colours he drew an identical small circle and marked eight equidistant points on each of the circles. To obtain the expansion effect, Whittle displayed the rows of colour shades so that the right eye would observe the row of red shades in the upper part of the visual field while the left eye observed the row of blue shades in the lower part of the visual field. Observed in binocular vision against a neutral background, these colour shades are simply experienced as rows of reddish and bluish shades. The crucial step in the experiment is the simultaneous change of the neutral background in the visual fields of the left and right eye, so that the background colour for the left eye is the reference colour in the centre of the blue shades and the background for the right eye is the reference colour for the red shades. Whittle's experiments showed that this has the effect that the two rows are experienced as two rainbows against a neutral background.

Naturally, the qualia compressions and expansions discussed in this section are all detectable – else we would not have known about them. In that respect, they differ from the type of compressions defined in Decock and Douven (2013), which are constructed in such a way as to make them *un*detectable (as explained in Sect. 4). Still, the foregoing shows that the idea of a qualia compression is not outlandish, given that such compressions appear to occur in reality. (The same holds true for the kind of translations that we consider in Fig. 1c. After all, it is hard to make sense of the results of Whittle's experiment without supposing that the centres of both the red and the blue shades have been shifted to the location that the shade of the neutral background occupies in phenomenal colour space.) And this in turn shows that the qualia compression scenarios from our earlier paper have, like some qualia inversion scenarios, a clear connection to science.

4 Qualia Change and Some Unresolved Issues in Colour Science

The foregoing considerations suggest that, at present at least, qualia compression is a scientifically plausible colour qualia change scenario. It escapes the major objection to qualia inversion – the asymmetry of phenomenal colour space – and contemporary colour science shows that qualia compression is not merely the product of philosophical speculation. Still, future developments in colour science might necessitate a different assessment of qualia compression scenarios, and perhaps also of colour inversion scenarios. In closing, we consider three directions that such developments might take and the repercussions these developments might have for the issue of qualia change.

First, it seems theoretically possible that phenomenal colour space is non-Euclidean. Indeed, in view of the difficulties involved in providing accurate Euclidean colour appearance models, it has been suggested that colour appearance models representing phenomenal colour space cannot be Euclidean. However, one must observe that colour appearance models are only graphical representations of phenomenal colour judgments obtained by statistical techniques such as multidimensional scaling. From a methodological point of view, it is hard to establish how faithfully such colour appearance models can represent phenomenal colour space. There is currently no positive evidence suggesting that colour appearance models should be non-Euclidean, and it is even doubtful that first person judgments, which are crucially involved in generating data concerning the geometrical structure of phenomenal colour experience, are reliable enough to ever warrant that conclusion. Still, it must be admitted that if phenomenal colour space could be shown to be non-

Euclidean, then that might well jeopardize qualia compression scenarios, given that it is far from obvious that an isomorphic miniature model of a space can be found within the space if that space is non-Euclidean.

Second, an interpretation of gamut expansions different from the one given in the previous section is possible. In the qualia compression argument, a *global* compression of phenomenal colour space is considered. The evidence from the experiments discussed above could also be interpreted as indicating that *local* compressions and expansions in phenomenal colour space are possible. Hence, one might suggest that phenomenal colour space has no stable metrical structure and that it should be interpreted as a topological space. But note that if there is no reason to consider *geometrical* isomorphisms in phenomenal colour space, we could consider instead *topological* isomorphisms, that is, homeomorphisms. Homeomorphisms in phenomenal colour space are much easier to construct than geometrical isomorphisms even for spaces that do have a metrical structure. In this view, the qualia change scenario is not jeopardized; rather, it is likely that in addition to the qualia compression scenario, the qualia inversion scenario could be saved. The alleged asymmetry of phenomenal colour space poses a problem for geometrical isomorphisms, but not for homeomorphisms. To put it graphically, the asymmetries in phenomenal colour space can be easily compensated for if one is allowed to squeeze and stretch phenomenal colour space for the purposes of defining a qualia inversion.

Third, some authors (e.g., Decock 2006) have suggested that there is no unique phenomenal colour space. In view of the fact that a plethora of shapes of phenomenal colour space has been presented in the past (see Kuehni 2003, pp 19–103), and that colour scientists use different colour appearance models for different viewing conditions, one could indeed doubt that there is a unique phenomenal structure that underlies all colour experience. On the view that neither metrical nor topological relations are stable through different viewing conditions and hence need not be preserved by a qualia change scenario, such scenarios are easy to come by. For then a qualia inversion scenario becomes a mere permutation of qualia. The qualia compression scenario would be a mapping of the (infinite) set of colour qualia into one of its subsets. In this view, neither qualia inversion nor qualia compression are problematic from a mathematical point of view, as there are no structural relations that are to be preserved.[7]

In summary, in view of the present scientific evidence, qualia compression is the most plausible qualia change scenario. Experimental results that might indicate that phenomenal colour space is merely a topological space or does not exist would not jeopardize the possibility of a large-scale qualia change. Only the discovery

[7] Clark (1985) also offers an argument that random qualia permutations are possible. He does not rely on the assumption that no structure between qualia need be preserved, but assumes instead that the structural relations between qualia can be reinterpreted at the same time, leaving all colour judgments invariant under the permutation of qualia and concomitant reinterpretation of the relations between qualia.

that phenomenological colour space is non-Euclidean might block all qualia change scenarios, but, as stated earlier, it is hard to see how that could be discovered even if it were true. Hence, although Locke himself thought lightly of it, the sceptical problem he raised has not been discarded so far, nor is it likely to undergo this fate any time soon.

Acknowledgement We are greatly indebted to an anonymous referee for valuable comments on a previous version.

References

Broackes, J. (2007). Black and white and the inverted spectrum. *Philosophical Quarterly, 57*, 161–175.

Brown, R., & MacLeod, D. (1997). Color appearance depends on the variance of surround colors. *Current Biology, 7*, 844–849.

Byrne, A. (2012). Inverted qualia. In *The Stanford encyclopedia of philosophy* (Winter 2012 Edition), http://plato.stanford.edu/entries/qualia-inverted/

Clark, A. (1985). Spectrum inversion and the color solid. *Southern Journal of Philosophy, 23*, 431–443.

Decock, L. (2006). A physicalist reinterpretation of 'phenomenal' spaces. *Phenomenology and the Cognitive Sciences, 5*, 197–225.

Decock, L., & Douven, I. (2013). Qualia compression. *Philosophy and Phenomenological Research, 87*, 129–150.

Ekroll, V., Faul, F., & Wendt, G. (2011). The strengths of simultaneous colour contrast and the gamut expansion effect correlate across observers: Evidence for a common mechanism. *Vision Research, 3*, 311–322.

Gilchrist, A. (2006). *Seeing black and white*. Oxford: Oxford University Press.

Hardin, C. (1988). *Color for philosophers*. Indianapolis: Hackett.

Jameson, D., & Hurvich, L. (1957). An opponent-process theory of color vision. *Psychological Review, 64*, 384–404.

Kuehni, R. (2003). *Color space and its divisions*. New York: Wiley.

Li, X., & Gilchrist, A. (1999). Relative area and relative luminance combine to anchor surface lightness values. *Perception and Psychophysics, 61*, 771–785.

Myin, E. (2001). Color and the duplication assumption. *Synthese, 129*, 61–77.

Quine, W. V. O. (1975). The nature of natural knowledge. In S. Guttenplan (Ed.), *Mind and language* (pp. 67–81). Oxford: Clarendon.

Whittle, P. (2003). Contrast colours. In R. Mausfeld & D. Heyer (Eds.), *Colour perception: Mind and the physical world* (pp. 115–138). Oxford: Oxford University Press.

Explanatory Coherence, Partial Truth and Diagnostic Validity in Psychiatry

Panagiotis Oulis

Abstract The paper deals with the thorny problem of the validity of psychiatric diagnostic concepts which engages core-issues in the philosophy of science such as those of truth and explanation. After an initial explication of the main facets of this concept in psychiatric diagnostic classification, I develop an account of psychiatric diagnostic concepts as conceptual models with clinical-descriptive and non-clinical explanatory parts and show the differential role they play in the justification of diagnostic concept-validity in the rest of medicine. Moreover, I elaborate comparative empirical and theoretical validity-criteria of psychiatric diagnostic concepts in terms of their partial truth. My account relies on the notions of explanatory coherence and mechanistic explanation. Furthermore, I apply more extensively this account to the specific example of the diagnostic validity of acute psychotic disorders. I conclude that both descriptive and explanatory considerations are necessary for the assessment and improvement of diagnostic validity in psychiatry.

1 Introduction

Among the foundational problems facing contemporary psychiatry, the problem of the validity of its diagnostic concepts, such as e.g. those of schizophrenic or bipolar disorders, remains not only unsolved but even very poorly understood. As a foundational problem, this problem engages several core-issues in the philosophy of science such as those of truth and explanation. In a generic and uncontroversial sense, "psychiatric diagnostic validity" means the extent to which psychiatric diagnostic concepts represent accurately distinctive patterns of human psychopathology. The currently prevailing diagnostic system in psychiatry is the Diagnostic and

P. Oulis (✉)
First Department of Psychiatry, School of Medicine, University of Athens, Athens, Greece
e-mail: oulisp@med.uoa.gr

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_35,

Statistical Manual of mental disorders (American Psychiatric Association 2000). This manual specifies explicit clinical criteria representing the clinical features of each diagnostic concept in the form of a long list of criteria. For the diagnosis of a specific mental disorder, a minimum number of these criteria are required (e.g. five out of nine for the diagnosis of an episode of major depression). The shortcomings of this manual include the excessive clinical heterogeneity of patients subsumed under the same diagnostic category and the substantial overlapping of the clinical features of allegedly distinct mental disorders. For example, patients with severe symptoms of major depression regularly experience also severe symptoms of generalized anxiety disorder and thus qualify for both diagnoses. This lack of sharp boundaries between distinct mental disorders has led two eminent scholars to claim that, at least for the time being, psychiatric diagnoses are devoid of any validity (Kendell and Jablensky 2003). Two major theses underlie their negative verdict: first, the thesis that validity is a categorical (all-or-none) property of psychiatric diagnostic concepts and, second, the thesis that only strictly descriptive considerations are necessary for the appraisal of their validity. A corollary of this thesis is that explanatory considerations are not necessary for the appraisal of psychiatric diagnostic validity. Space limitations do not allow here my critical engagement with this very influential in contemporary psychiatry paper (cited already almost 600 times as per Google Scholar, July 2012). However, the alternative account I will develop aims to show that diagnostic validity in psychiatry is a matter of degree and that explanatory considerations are also necessary for the appraisal of psychiatric diagnostic validity. In the following, I first explicate briefly the main concepts of validity involved in psychiatric diagnostic classification. Then, I develop an account of psychiatric diagnostic concepts as conceptual models with clinical descriptive and non-clinical explanatory parts of three main types. I explicate further these component-parts as well as the differential role they play in the justification of diagnostic concept-validity in the rest of medicine. Moreover, I elaborate comparative empirical and theoretical validity-criteria of psychiatric diagnostic concepts in terms of their partial truth. My account relies on the notions of explanatory coherence and mechanistic explanation. Furthermore, I apply more extensively this account to the specific example of the diagnostic validity of the concept of acute psychotic disorders. This case-study serves as a test-case of the adequacy of my account. Finally, I conclude that both descriptive and explanatory considerations are necessary for the appraisal of diagnostic validity in psychiatry.

2 On the Concepts of Validity in Psychiatric Diagnosis

In a generic and uncontroversial sense, "validity" of psychiatric diagnostic concepts means whether and to what extent they represent accurately distinct patterns in the three main components of mental disorders: clinical, pathophysiological and aetiopathogenetic. The clinical component includes patients' patterns of abnormal experiences and behaviors (clinical symptoms and/or signs). These patterns of co-occurrence of these clinical symptoms and signs, along with their characteristic clinical course, are called *clinical syndromes* and the type of validity of their respective diagnostic concepts *clinical validity*. Furthermore, their pathophysiological component subsumes the patterns of factors and mechanisms underlying clinical symptom and/or sign formation of each clinical syndrome (*pathophysiological validity*). Finally, their aetiopathogenetic component subsumes the patterns of factors and mechanisms underlying patients' latent predisposition or strong vulnerability to the eventual development of the clinical syndrome of a mental disorder (*aetiopathogenetic validity*). Jointly, clinical, pathophysiological and aetiopathogenetic validity constitute *diagnostic validity*. In the assessment of diagnostic validity several other concepts of validity, borrowed from psychometric theory, are invoked as well. These concepts of validity, spanning the three components of mental disorders (clinical, pathophysiological and aetiopathogenetic), revolve around *external validity* and its several logical subtypes such as antecedent, concurrent and predictive validity. More precisely, "*antecedent*" *validity* means the strength of associations of the clinical syndrome with factors operating long before the development of the clinical manifestations of the disorder (e.g. family history for similar mental disorders, premorbid personality features, early adverse life-events etc). "*Concurrent*" *validity* means the strength of associations of the clinical syndrome with several concomitant features (e.g. findings of brain imaging scans, performance in neuropsychological tests, recent adverse life-events, performance-level in major social roles, etc) and finally, "*predictive*" *validity,* the strength of associations of the clinical syndrome with future outcomes (e.g. favorable response to specific treatments, duration and quality of recovery, etc) (see, e.g. Kendler 1990; Schaffner 2012).

3 Psycho-Diagnostic Concepts as Psychopathological Models

Current psychiatric diagnostic concepts are hypothetical clinical conceptual models of presumably homogeneous classes of real mental patients with respect to their clinical symptoms and/or signs and regularities. In other words, the sets of diagnostic criteria of these concepts sketch *minimal* conceptual models of their respective hypothesized mental disorders. These models are epistemologically abstract since they refer directly to ideal or typical patients of these classes and only indirectly to their real members in the population at large. Moreover, they are methodologically abstract in that they represent only the most salient or characteristic co-occurring clinical features of the hypothesized disorders, omitting deliberately a host of potentially relevant patient-features. A similar distinction holds with respect to their truth. Whereas they are true of the members of their direct reference classes, they are at best only partially true of the members of the indirect ones.

I said previously that psychiatric diagnostic concepts as minimal conceptual models represent in their explicit diagnostic criteria only the more salient clinical features and regularities of their respective hypothesized disorders. However, psychiatric researchers investigate, in a hypothesis-driven manner and with various degrees of success, a host of further psychopathological features associated with the diagnostic ones. These associations come in the form of law-like ceteris paribus generalizations of the following main types (for a more detailed classification and analysis of psychopathological generalizations, see Oulis 2013):

1. *Associated* clinical generalizations, representing further clinical features not represented in their diagnostic criteria. Examples: treatment-response, social and occupational functioning, long-term outcome etc.
2. *Pathophysiological* generalizations, representing factors leading to the emergence of the diagnostic clinical features of the disorder. Examples: neurobiological or neuropsychological abnormal changes, acute intoxication with psychoactive substances, acute exposure to severe triggering adverse life-events etc.
3. *Aetiopathogenetic* generalizations, representing factors underlying the development of the clinically latent pathological basis or the strong vulnerability to the disorder. Examples: predisposing genetic factors, early environmental adversities, pre-morbid temperament and personality, chronic psycho-social stress etc.
4. *Mixed* generalizations of the previous types.

The set of the previous types of generalizations, jointly with their diagnostic criteria, constitute *expanded,* though far from complete, conceptual models of mental disorders. In other words, psychiatric diagnostic concepts as expanded psychopathological models consist of the *whole* cluster of their diagnostic and associated generalizations. Correlatively, whereas their clinical parts represent only "nominal essences", their expanded models intend to represent their "real essences".

Diagnostic validity in medicine involves the same types of generalizations, however with differential disease-status. To begin with, the main diagnostic clinical features along with their associated clinical features and patterns constitute *clinical syndromes* or *disease-families*. Example: polyuria, polydipsia and polyphagia, along with their clinical complications constitute the clinical syndrome of diabetes mellitus. Most psychiatric diagnostic concepts of DSM-IV have still this disease-status. This is precisely why the current debate on their degree of validity focuses on whether their characteristic pattern of symptoms/signs exhibits sufficient stability, uniformity and discrete clinical boundaries from the clinical syndromes of related mental disorders. Strong clinical validity enables accurate prediction of clinical course and outcome. Moreover, it might also help predict patients' response to available treatments, if any.

Furthermore, clinical syndromes jointly to their associated pathophysiological features constitute *disease-genera*. Example: the clinical syndrome of diabetes mellitus and high fasting blood glucose levels constitute diabetes mellitus as a generic disease. Finally, clinical syndromes, along with their main pathophysiological and aetiopathogenetic features individuate *disease-species*. Example: generic diabetes mellitus and pancreatic failure to produce insulin constitute type I-diabetes, whereas generic diabetes and body-tissue resistance to insulin constitute diabetes of type II.

Strong validity of psychiatric diagnostic concepts as expanded psychopathological models, presupposes strong empirical confirmation of their component-generalizations. Moreover, the normative epistemological relationships between these two parts of psychopathological expanded models are these: the pathophysiological and/or aetio-pathogenetic parts should be explanatory of the clinical, especially the diagnostic ones, whereas the latter should provide a decisive testing ground of the validity of the former. Thus, strong empirical partial truth and explanatory or theoretical partial truth jointly ground the validity of psychiatric diagnostic concepts as expanded models. I now turn to a more precise characterization of the major indicators of empirical and theoretical partial truth in the field of psychiatric diagnostic research.

4 Empirical and Theoretical Partial Truth-Indicators of Expanded Psychopathological Models

4.1 Empirical Partial Truth-Indicators

Both minimal and expanded psychopathological models are composed of law-like psychopathological generalizations which are *ceteris paribus* ones. However, they are empirically testable in principle, owing to the randomization-methods of research sample-selection. Indeed, these methods allow-within the limits of random sampling error-for the control of all confounding factors, both known and unknown, collectively represented in their ceteris paribus clauses. Moreover, psychopathological law-like generalizations are mostly *comparative* or *contrastive* ones, displaying the magnitude of qualitative or quantitative differences between experimental and control groups. The typical pattern of qualitative psychopathological generalizations is this: "Ceteris paribus, patients with mental disorder x display more often than controls feature y with a frequency z%". Likewise, the typical pattern of quantitative psychopathological generalizations is this: "Ceteris paribus, patients with mental disorder x display more of feature y than controls by an amount of z%".

Meta-analytic studies assess in a systematic manner the consistency and the magnitude of these differences, technically called "effect-sizes". In bivariate group comparisons, the most frequently used measures of effect-size are the odds-ratio for binary variables and the Cohen's d or Hedges g for continuous variables. Effect-sizes are further normatively evaluated as weak, moderate or strong and can thus help assess the empirical partial truth of psychopathological generalizations (e.g. d-values of 0.2–0.4 are considered as weak, 0.5–0.7 as medium and >0.8 as strong). More precisely, psychopathological generalizations as components of expanded psychopathological models with consistently strong effect-sizes could be considered as approximately true. It is tempting then to take the ratio of mutually independent approximately true generalizations (n) to the total number of mutually independent generalizations of the expanded conceptual models (N) as an indicator of their overall empirical partial truth (*e*=*n*/*N*). However, this proposal faces several problems. Some of these problems are the following:

(a) Although sufficient for the assessment of the degree of clinical validity of psychiatric diagnostic concepts as minimal models, this proposal is clearly insufficient for the assessment of their degree of validity as expanded models. This is so because the latter contain non-observational concepts figuring in non-observational generalizations which are empirically tested only through some of their deductive consequences in conjunction with further auxiliary assumptions and empirical data relevant to them. This is the well-known Duhem-Quine or the empirical under-determination problem.
(b) The proposal leaves out of account the differential evidential weight of explanatory generalizations' projective performance, especially in making successful risky predictions.
(c) Expanded psychopathological models are usually only rough sketches of their respective mental disorders. Moreover, their component-generalizations are often vague and only loosely inter-connected.

Thus, this proposal provides at best only a partial elucidation of the concept of empirical partial truth of expanded psychopathological models, rather than a strict measure thereof. At any rate, this proposal, along with the following ones are only intended to help assess the diagnostic validity of alternative models of the same mental disorder, i.e. in a *comparative* manner.

4.2 Theoretical Partial Truth-Indicators

In addition to the requirement of strong empirical confirmation or coherence with empirical data, pathophysiological and aetiopathogenetic generalizations of expanded psychopathological models should also be explanatory of the diagnostic and associated clinical generalizations. Moreover, they should form a web of mutually logically consistent and explanatorily coherent generalizations. Their explanatory coherence is strengthened by mechanistic explanations. Mechanism-based explanations, disclosing the underlying pathophysiological and/or aetiopathogenetic mechanisms of clinical diagnostic and associated features and generalizations are pervasive in the whole of medicine (for an analysis of the nature of psychopathological mechanisms, their main types and their relevance to psychiatric diagnostic classification, see Oulis 2010). Among contemporary philosophers of science, Paul Thagard has particularly stressed the importance of increased explanatory coherence through successful mechanistic explanations at increasingly deeper levels for resisting the "pessimistic induction" argument against scientific realism (Thagard 2007). One can doubt whether this requirement is necessary for resisting pessimistic induction. Still, one cannot doubt that independently confirmed mechanistic explanations of previously posited unobservable entities and/or their properties (and changes thereof, that is, events and processes), provide crucial though not infallible theoretical evidence for the genuine existence and thus for the approximate truth of scientific hypotheses positing them. Indeed, the best explanation of their independent strong empirical confirmation is that they are at least approximately true. In the following, I will illustrate the relevance of explanatory coherence through deep mechanistic explanations to the validation of expanded psychopathological models in the specific example of acute psychotic disorders and their diagnostic demarcation from schizophrenic disorders. This case-study will serve as a test-case for the adequacy of my general account.

5 Acute Psychotic Disorders as a Case-Study

Acute psychotic disorders are a still officially non-recognized clinical syndrome with substantially overlapping clinical boundaries with the otherwise much broader clinical syndrome of schizophrenia. Acute psychotic disorders include the DSM-IV diagnostic categories of brief psychotic disorder and schizophreniform disorder which share the same diagnostic symptoms and/or signs, with the sole exception of their total duration (up to 1month and 6 months, respectively) (APA 2000). As a theoretical model of schizophrenia the model I will describe is partial in that it intends to account only for a part of the clinical syndrome of schizophrenia and only for its acute onset and its subsequent acute relapses. In particular, it does not apply to the whole clinical syndrome of schizophrenia at all stages of its clinical course as a presumably unitary chronic psychotic disorder, for which the validity of the dopamine hypothesis I will present in the following is highly doubtful (see Kendler and Schaffner 2011). Delusions, often bizarre, along with auditory hallucinations and disorganized thinking and behavior are the cardinal clinical features of acute psychotic disorders. Delusions are firmly sustained beliefs about internal or external reality in relation to the patients themselves, involving their basic physical and/or psychological needs, such as those for safety, dignity, or worthiness, which are held with absolute conviction despite overwhelming evidence to their falsity and even blatant irrationality. For example, a male patient on waking up in the morning and feeling pain in the abdomen, becomes firmly convinced that overnight he has been transformed into a female pregnant woman by god and will soon deliver the new messiah, or another that a dead bird found on the street is an god-sent anticipatory sign of the impending end of the world.

It has been known since decades that drugs increasing dopamine neural-transmission in the brain may cause acute schizophrenia-like psychoses, as well as that drugs blocking central dopaminergic neuro-transmission are efficacious in their treatment. Later on, it has been discovered that anti-psychotic drugs blocked preferentially the dopamine-type 2 receptors in the mesolimbic dopamine pathway of the brain (a pathway connecting the ventral tegmental area in the midbrain to the nucleus accumbens in the limbic system) and, more recently, that the brain-striatum of patients (a core-component of basal ganglia, a group of nuclei at the base of the forebrain) has a greater dopamine-synthesis capacity than that of normal controls, with a substantially strong effect-size (Howes et al. 2011). All the preceding well-confirmed mixed causal-probabilistic ceteris paribus generalizations constituted more precise specifications of the explanatory role of elevated dopamine in the pathophysiology of acute schizophrenia-like psychoses. However, they still did not specify any mechanism through which elevated dopamine synthesis leads to delusion-formation. In other words, they did not specify any precise and empirically testable pathophysiological mechanism explanatory of the causal role of dopamine in delusion-formation. A fortiori, they did not explain the aetiopathogenetic mechanisms of patients' pre-morbid strong propensity to increased dopamine release in their brain-striatum.

More recently, it has been proposed one such mechanistic hypothesis which currently undergoes indirect empirical clinical testing with encouraging preliminary results (see for recent reviews Howes and Kapur 2009; Heinz and Schlagenhauf 2010). This hypothesis constitutes the core of an expanded but still only partial model of acute schizophrenic disorders (called "the aberrant salience model") since it intends to account for the final only causal path (pathophysiological) of only some of their cardinal clinical features (especially delusions). However, it coheres explanatorily with several successfully empirically tested aetiopathogenetic causal-probabilistic hypotheses. More precisely, the aetiopathogenetic background of the aberrant salience hypothesis is this: Multiple heterogeneous causal-probabilistic factors, exerting their influence through still only partially understood pathogenetic mechanisms, underlie individuals' increased vulnerability to acute psychoses. These empirically well-confirmed risk-factors include genes, paternal age at conception, maternal infections during pregnancy, obstetric complications at delivery, early chronic cannabis abuse, and several markers of social adversity such as migration, unemployment, urban upbringing in extreme poverty, childhood abuse, social isolation and lack of close friends. Each of these factors increases individuals' risk for a schizophrenia-like psychotic illness by two- to almost fifth-fold (MacDonald and Schulz 2009). Their common final effect, grounding individuals' increased vulnerability to acute psychotic disorders, consists in the sensitization of patients' corpus striatum (a relay-station of the mesolimbic dopamine pathway), that is in its acquisition of a strong propensity to elevated pre-synaptic dopamine synthesis and release in the synaptic cleft (see e.g. Broome et al. 2005). In turn, increased dopamine-release in the corpus striatum confers strong personal significance ("motivational salience") to our moment-by-moment perceptual experiences, whether of external or internal stimuli. This mechanism is presumably rooted in our evolutionary past and has been selected because of its positive contribution to individual organisms' survival (e.g. perceptual experiences signaling danger lead to prompt avoidance of its source). Moreover, this mechanism underlies both classical and operant associative learning and is activated in response to mismatches between actual and predicted events (technically called mechanism of "prediction error"). The specific pathophysiological mechanistic hypothesis is then this: In individuals with increased vulnerability to acute psychoses, this mechanism fails and becomes activated by internal or external perceptual experiences devoid of any objective vital importance for the subject. More precisely, this mechanism of elevated striatal dopamine synthesis and release is activated in patients in the absence of any mismatch between expectations and outcomes, that is, in the face of predictable or neutral events (Morris et al. 2012). This activation accounts for the enigmatic but vital importance of otherwise trivial and innocuous perceptual experiences evoking in patients strong feelings of perplexity as well as the subjectively pressing need to explain their importance to themselves. Underlying this psychological process, associative striatal regions activate cortical association areas carrying out the psychological function of thinking. Eventually, patients' psychological process of explanation-seeking for their abnormally salient perceptual experiences culminates in delusion-formation, however bizarre these might be.

In the previous example, the overall partial truth of the dopamine hypothesis of acute psychotic disorders is strengthened by its explanatory integration with the aberrant salience hypothesis specifying the main mechanism mediating between elevated striatal pre-synaptic dopamine synthesis and release and delusion-formation. Moreover, the best explanation of the independent experimental confirmation of the aberrant salience mechanistic hypothesis is the approximate truth of the dopamine hypothesis. Furthermore, its approximate truth is further strengthened by its progressive explanatory integration with the whole cluster of independently empirically confirmed aetiopathogenetic causal–probabilistic ceteris paribus generalizations about acute psychotic disorders. Finally, it is consistent and can be further explanatorily integrated with our knowledge about the evolutionary origins of the dopamine-salience functional system.

More precisely, the core-components of this partial expanded model are the following:

(a) The diagnostic features of acute psychotic disorders, the clinical hypothesis of their regular co-occurrence as a distinct clinical syndrome and their associated ceteris paribus generalizations, including their favorable response to dopamine-type 2 receptors-blocking pharmacological agents (clinical domain). The latter generalization suggests as very likely the central role of dopamine in the generation of acute psychotic symptoms, without however explaining mechanistically this role.

(b) The pathophysiological hypothesis of the brain-striatum dopamine-mediated aberrant personal salience of internal or external perceptual experiences as a mechanism of acute delusion-formation in subjects with prior sensitization of their brain-striatum and correlatively strong propensity to elevated pre-synaptic dopamine- synthesis and release (pathophysiological domain).

(c) The cluster of aetiopathogenetic ceteris paribus generalizations about the biological and psycho-social factors leading to the development of the underlying strong vulnerability to psychotic disorders, that is to the acquisition of the lasting propensity of brain-striatum to elevated pre-synaptic dopamine-synthesis and release (aetiopathogenetic domain).

The aberrant salience partial model integrates explanatorily the clinical and therapeutic generalizations about acute psychotic disorders. In so doing, it strengthens rational belief in the previously posited causal role of dopamine in delusion-formation. More precisely, the strong empirical partial truth enjoyed by this hypothesis on the grounds of mere clinical pharmacological considerations is now supplemented by a non-negligible theoretical partial truth as well, owing to the explanatory mechanistic depth of the aberrant salience hypothesis. Further still, the aberrant salience hypothesis now progressively integrates explanatorily the whole cluster of empirically confirmed aetiopathogenetic generalizations of acute psychotic disorders, by exhibiting the main mechanism underlying their strong law-like associations. Thus, it unifies explanatorily a host of disparate robust findings from several research-fields ranging from genetics to social epidemiology. By the same token, the explanatory parsimony of the expanded psychopathological model of acute psychotic disorders is greatly enhanced as well. Moreover and conversely, the increasing explanatory coherence of the theoretical part of the aberrant salience model strengthens in turn the clinical validity of the concept of acute psychotic disorders as representing a *distinct* clinical syndrome and provides crucial theoretical evidential support for its clinical demarcation from the presumably unitary diagnostic concept of schizophrenia. However, the clinical explanatory range of the aberrant salience model remains still limited since it accounts for only one out of the three cardinal clinical features of acute schizophrenic disorders, though some leading researchers suggest that it explains auditory hallucinations as well (e.g. Kapur 2003).

Acute psychotic disorders overlap clinically with schizophrenic disorders. Thus, the aberrant salience expanded model of acute psychotic disorders is an only *partial* expanded model of schizophrenia. More comprehensive expanded models of the diagnostic concept of schizophrenia should be built by *explanatory integration* of this partial model with further partial models of the remaining parts of its clinical syndrome, e.g. of its so-called negative clinical signs, including affective flattening, poverty of speech and inability to initiate and persist in goal-directed activities. Moreover, they should integrate several robust research-findings relevant to schizophrenia from the fields of social epidemiology, genetics, neuroanatomy, neuro-physiology, and neuropsychology (see e.g. for a review MacDonald and Schulz 2009). This integration should be the outcome of a process of mutual adjustments of co-evolving partial models of schizophrenic disorders, including drastic revisions of their current minimal psycho-diagnostic model, i.e. of their current diagnostic criteria. The aim of this integration should be the joint optimization of their empirical and theoretical partial truth, assessed at least in a *comparative* manner between alternative models of the same mental disorder. In general, psychiatric diagnostic research aiming at the discovery of disease-genera is more realistic than the search for disease-species. This is so because of the closer proximity of pathophysiological factors and/or mechanisms to patients' clinical symptoms/signs by comparison with their far more distal aetiopathogenetic ones. This strategy contrasts with the persistent attempts to ground psychiatric diagnostic classification almost exclusively in distal genetic or epigenetic factors (see e.g. in the case of schizophrenia Craddock and Owen 2005 and Crow 2008, respectively).

6 Conclusions

The validity of psychiatric diagnostic concepts as tentative models of human psychopathology comes in degrees and can be assessed by the proportion of their partially true component generalizations in at least a comparative manner. Moreover, diagnostic concepts in both medicine and psychiatry can be differentially valid as disease- families (clinical syndromes having at most strong clinical validity only), disease-genera (enjoying strong clinical and pathophysiological validity) or disease-species (with strong clinical, pathophysiological and aetiopathogenetic validity). Valid concepts of disease-families or clinical syndromes have predictive but no explanatory power, whereas valid concepts of disease-genera and disease-species have both. Strong explanatory coherence and integration of the component-parts of expanded models of mental disorders through successful mechanistic explanations of increasing depth is a crucial theoretical indicator of their validity as disease-genera or species, again as assessed in a comparative way between alternative expanded models of the same mental disorder. At present, besides their clinical heterogeneity, psychiatric diagnostic concepts as expanded models of mental disorders are also heterogeneous with respect to both their pathophysiological and aetio-pathogenetic mechanisms as well. The discovery and incorporation in their expanded models of distinct central pathophysiological mechanisms leading to otherwise similar clinical syndromes would increase their pathophysiological validity and allow their more valid diagnostic identification as *distinct* clinical-pathophysiological syndromes or psychopathological genera. Moreover, this would in turn enable a more refined and accurate *re-description* of their initially unitary and undifferentiated clinical syndromes, with considerable improvements of the clinical validity of their refined successor-clinical syndromes. Thus, explanatory considerations are also necessary, along with merely descriptive ones, for both the assessment and improvement of psychiatric diagnostic validity. I have tried to illustrate extensively the explanatory coherence approach to psychiatric diagnostic validity in the example of the aberrant salience theoretical model of acute psychotic disorders and its decisive role in their diagnostic demarcation from chronic schizophrenia. As this example suggests, deeper explanations through independently confirmed mechanistic hypotheses and stronger explanatory integration of expanded psychopathological models provide decisive theoretical evidence for the comparative assessment and improvement of psychiatric diagnostic validity.

Acknowledgments Many thanks to Stathis Psillos and an anonymous reviewer for their helpful comments.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders, fourth edition, text revision (DSM-IV-TR).* Washington: American Psychiatric Association.

Broome, M. R., Woolley, J. B., Tabraham, P., Jojns, L. C., Bramon, E., Murray, G. K., Pariante, C., McGuire, P. K., & Murray, R. M. (2005). What causes the onset of psychosis? *Schizophrenia Research, 79*, 23–34.

Craddock, N. J., & Owen, M. J. (2005). The beginning of the end for the Kraepelian dichotomy. *British Journal of Psychiatry, 186*, 364–366.

Crow, T. J. (2008). Craddock and Owen versus Kraepelin: 85 years late, mesmerized by "polygenes". *Schizophrenia Research, 103*, 156–160.

Heinz, A., & Schlagenhauf, F. (2010). Dopaminergic dysfunction in schizophrenia: Salience attribution revisited. *Schizophrenia Bulletin, 36*, 549–562.

Howes, O., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: Version III-the final common pathway. *Schizophrenia Bulletin, 35*, 549–562.

Howes, O. D., Bose, S. K., Turkheimer, F., Valli, I., Egerton, A., Valmaggia, L. R., Murray, R. M., & McGuire, P. (2011). Dopamine synthesis capacity before onset of psychosis: A prospective [18F]-DOPA PET imaging study. *American Journal of Psychiatry, 168*, 1311–1317.

Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology and pharmacology in schizophrenia. *American Journal of Psychiatry, 160*, 13–23.

Kendell, R. E., & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry, 160*, 4–12.

Kendler, K. S. (1990). Towards a scientific psychiatric nosology: Strengths and limitations. *Archives of General Psychiatry, 47*, 969–973.

Kendler, K. S., & Schaffner, K. F. (2011). The dopamine hypothesis of schizophrenia: An historical and philosophical analysis. *Philosophy, Psychiatry and Psychology, 18*, 41–63.

MacDonald, A. W., & Schulz, S. C. (2009). What we know: Findings that every theory of schizophrenia should explain. *Schizophrenia Bulletin, 35*, 493–508.

Morris, R. W., Vercammen, A., Lenroot, R., Moore, L., Langton, J. M., Short, B., Kulkarni, J., Curtis, J., O'Donnell, M., Weickert, C. S., & Weickert, T. W. (2012). Disambiguating ventral striatum fMRI-related bold signal during reward prediction in schizophrenia. *Molecular Psychiatry, 17*, 280–289.

Oulis, P. (2010). Nature and main kinds of psychopathological mechanisms. *Dialogues in Philosophy, Mental and Neuro Sciences, 3*, 27–34.

Oulis, P. (2013). Towards a unified framework for the science and practice of integral psychiatry. *Philosophy, Psychiatry and Psychology, 20*, 113–126.

Schaffner, K. F. (2012). A philosophical overview of the problems of validity for psychiatric disorders. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 169–186). Oxford: Oxford University Press.

Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science, 74*, 28–47.

Part XII
Philosophy of the Social Sciences

Performativity: Saving Austin from MacKenzie

Uskali Mäki

Abstract The new economic sociology claims to have adopted the notion of performativity from J.L Austin, has put it in new uses, and has given it new meanings. This is now spreading and has created another vogue term in the social and human sciences. The term is taken to cover all sorts of aspects in the ways in which the use of social scientific theories have consequences for the social world. The paper argues that the expansive use of ‘performativity’ obscures the Austinian idea and thereby impoverishes the conceptual resources available for analyzing the nuances in the complex theory/world connections. Importantly, it blurs the difference between constitutive and causal relationships, both of which actually are involved. Instead of *economics performs the economy* as the sociologists say, it would make more sense to say, *the economy performs economics* – but even this would be undermotivated.

1 Introduction

Performativity is a new vogue word in the vocabulary of contemporary social science. Next to its other instantiations, sociologists Michael Callon, Donald MacKenzie and others have argued that economics has a “performative” relationship with the economy. Economics does not describe and explain a pre-existing economy, but rather shapes the social world by “performing” it. For example, the structure of financial markets and the practice of finance are influenced by modern finance

U. Mäki (✉)
Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences,
University of Helsinki, Helsinki, Finland
e-mail: uskali.maki@helsinki.fi

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_36,

theory: the latter "performs" the former (MacKenzie 2006a, b). For another, economists have been active in advising governments in Bolivia, Chile, Poland and Russia by designing markets and policies: again, the latter are being "performed" by economics (MacKenzie et al. 2007, p. 2). This way of speaking has now become popular within the social sciences more broadly when characterizing the ways in which social scientific theory and research relate to the social world. The insight is said to be "the most challenging recent theoretical contribution to economic sociology" (MacKenzie and Millo 2003, p. 107).

These new performativity theorists have adopted the term from J.L. Austin's theory of speech acts, apparently believing that this will bring illumination to the intricate ways in which economic ideas and practices are intermingled. I remain unconvinced about this and will try to point out that instead of illuminating, this terminology has managed to obscure an important set of facts about social reality. The term had better be returned and restricted to its original use. (For another set of queries, see Didier 2007.)

My focus will be on terminological and conceptual issues, so I leave aside the empirical issue of whether the claims of the new sociological performativists are supported by empirical evidence. Like the earlier and similar idea of "social construction" and its kin, "performativity" has remained unclear in its precise contents and consequences. Expanding on earlier work (Mäki 2008, 2012), the paper briefly examines the notion, spelling out in some detail why the relationship between economics and the economy is not Austinian-performative, and arguing that there is no reason to obscure the Austinian notion by extending the domain of 'performativity' far beyond its authentic Austinian domain; this only leads to an impoverishment of the conceptual resources available for recognizing the full diversity of aspects in the relationship between economics and the economy. Among the advocates of the performativity thesis, my main focus is on MacKenzie's formulations since they tend to be relatively more scrutable.

2 From Austin to MacKenzie

In his articles, books and talks, MacKenzie has been explicit in appealing to J.L. Austin's ideas. In a 2004 article (that promises to provide "conceptual clarification") MacKenzie starts putting forth a typology of two kinds of performativity. He draws a distinction between "Generic" and "Austinian" performativity. He explains the meaning of the latter thus: "To ask whether a model in financial economics is performative in the Austinian sense is to ask … whether the effect of the practical use of the model is to change patterns of prices towards greater compliance with the model" (2004, p. 306). MacKenzie has later augmented this typology, but before discussing the new one, let us consider some general characterizations.

It is hard to find a clear unambiguous definition of the notion of performativity in the sociological literature. It might be illuminating to consider the available characterizations as exemplifications of this simple general form of statements about performativity:

X performs *Y*

In the relevant new performativist literature, *X* is variously taken to denote things such as 'economics', 'economists' and 'financial models'; *Y* stands for 'markets' and 'economic processes' and 'economic relationships' and so on; while the relationship between *X* and *Y*, that of performing, is also referred to as 'shaping' and 'making' and 'constructing' etc. (There are other exemplifications, some of them rather confusing, but I put them aside here.) These expressions appear in the titles and subtitles of two representative books: *Do Economists Make Markets?* (MacKenzie et al. 2007) and *How Financial Models Shape Markets* (MacKenzie 2006b). So it would seem that generally, when *X* performs *Y*, it is the case that somehow *X* contributes to the existence or emergence or (change in the) properties of *Y*.

The "performativity thesis" is often contrasted with the idea that economic theory *describes and explains* economic phenomena. The new sociological performativists say they reject "traditional" views about science according to which science seeks to truly describe and explain phenomena. So goes Callon in characterizing what he disputes: "The discovery of formulas such as that of Ramsey or of Black-Scholes does not change behavior; it describes and clarifies it, just as Newton's laws have not changed the behavior of falling apples" (Callon 2007, p. 314). Disputing this idea looks similar to Austin's key point that performatives do not report or describe, truly or not, the speaker's actions; performatives help create new things rather than describe pre-existing things. According to Callon and MacKenzie, performing involves *changing rather than describing the world*. Callon puts this also by saying that "discourse *acts on* its object" (2007, p. 316), while MacKenzie talks about "option theory's practical consequences" (2006b, p. 6) and, employing a characteristic Austinian phrase, to "claim that economics is performative is to argue that *it does things*, rather than simply describing (with greater or lesser degrees of accuracy) an external reality that is not affected by economics" (MacKenzie 2006a, p. 29; italics added). Adapting the title of Austin's major book (*How to Do Things with Words*), Francesco Guala chooses to entitle his paper as "How to do things with experimental economics" (Guala 2007).

A further idea is that rather than describing and explaining the economy, watching it from outside as it were, economics changes it from within. Economics and economists are *inside rather than outside the economy*. "By participating in the economy, [economics] would place itself within the object it is supposed to be studying form the outside…" (Callon 2007, p. 315). MacKenzie puts it similarly: "The academic discipline of economics does not always stand outside the economy, analyzing it as an external thing; sometimes it is an intrinsic part of economic processes. Let us call the claim that economics plays the latter role *the performativity of economics*" (MacKenzie 2006b, p. 16).

So, when saying ‘*X* performs *Y*’ the new sociological performativists seem to be saying that ‘*X* changes *Y* from within Y’. But we don’t yet know what it is for *X* to change *Y*, and what it is for *X* to be within *Y*. We need to look elsewhere for further clarity. MacKenzie’s much-cited augmented typology of kinds of performativity (MacKenzie 2006b, p. 17, 2007, pp. 55–56) might be expected to offer some help:

> *Generic performativity*: An aspect of economics (a theory, model, concept, procedure, data set, etc.) is used by participants in economic processes, regulators, etc.
>
> *Effective performativity*: The practical use of an aspect of economics has an effect on economic processes.
>
> *Barnesian performativity*: The practical use of an aspect of economics makes economic processes more like their depiction by economics (while *counterperformativity* makes them look less like their depiction by economics).

As we will see in subsequent sections, these formulations are helpful for seeing the difference of these notions from the authentic Austinian notion of performativity. Here I want to make a few immediate observations. First, the typology suggests that it is not correct after all to take performativity generally to imply *making a contribution to a change* or *having a consequence*. This is a characteristic of effective and Barnesian versions only, but generic performativity lacks it – it only talks about economics *being used*. Second, the definition of Barnesian performativity (as well as counterperformativity) suggests that performativity, after all, does *not rule out the possibility of true description*. Indeed, the formulation implies that possibility. So it is not the case that if a theory “performs” its target, it therefore is not or cannot be (more or less) true about it.

Third, this new typology drops the attribute ‘Austinian’ and replaces it with ‘Barnesian’ (after the sociologist of scientific knowledge Barry Barnes), but the definition remains intact, which is to say that the same concept only becomes renamed. MacKenzie motivates this by saying that “the invocation of Austin could be read as suggesting that the performativity of economics was a linguistic matter” (MacKenzie 2006a, 29fn) and that to “analyze performative utterances using only linguistic philosophy is … to treat them as ‘magic’” (Mackenzie 2006a, p. 43, 2007, p. 68). He seems to suggest that in order to avoid treating performativity as merely linguistic magic, we need to see that their “felicity conditions” (conditions required for a performative to be effective) are social conditions. In the case of the wide practical adoption of the Black-Scholes-Merton model in the financial markets, these conditions have included the authority of economics, the model’s cognitive simplicity, and its public (also technical) availability (2006a, pp. 43–44, 2007, pp. 69–71). However, it is not clear why the idea of felicity conditions being social conditions would justify the move from ‘Austinian’ to ‘Barnesian’ given that Austin’s own open-ended list of felicity conditions includes requiring that right words be uttered by individuals with appropriate statuses in right circumstances, and that these things are governed by social conventions (Lectures II & III in Austin 1962).

Fourth, there *is* a very good reason for removing 'Austinian' from the typology as articulated by MacKenzie. It is the same reason that suggests removing the general umbrella term 'performativity'. Let me explain.

3 Constitution and Causation

On Austin's (1962) account of performativity, one performs an action by uttering some string of words, a performative sentence. If I say "I promise to deliver the paper by the deadline" I am thereby promising to deliver the paper by the deadline. To utter a performative sentence is not to *describe* a pre-existing action (that of promising), nor does "I promise" *cause* the promise. To utter the sentence *is* to promise, it is to perform the very action of promising. *Saying so makes it so* (provided the felicity conditions are met).

A key distinction in my argument can now be spelled out. The connection between speaking words and doing things is one of *constitution* rather than causation. Saying "I apologize" constitutes the act of apologizing. Saying "I agree" constitutes the act of agreeing. Those utterings do not *cause* those acts, rather those acts are *constituted* by those utterings. To utter those sentences *is* to take those actions.

This authentic meaning of performativity has been obscured by the literature on how economic theory can have consequences for economic reality. MacKenzie correctly recognizes the Austinian use of the term in characterizing certain speech acts in the world of finance such as when agreements and contracts are made. When, in response to an offer to sell or buy an asset at a particular price, someone says "done" or "agreed", then a deal is agreed (MacKenzie 2006b, p. 16). Indeed, uttering such words performs the act of agreement, that is, constitutes the act in a non-causal manner. This is genuine performativity. MacKenzie should have left it here.

However, right thereafter (2006b, pp. 17–19), and without warning or motivation, the extension of the word 'performativity' is vastly enlarged by offering the typology of three meanings that as such seem unrelated to the authentic meaning. These are the ones we cited above: generic performativity, effective performativity, and Barnesian performativity.

In none of these three types of case is the relationship between an aspect of economics and some aspect of the economy constitutive. A constitutive relationship would require that uttering or writing down an economic model for an audience (that understands the model and perceives the uttering as genuine and done in appropriate circumstances) establishes the model world as part of the real world. What is important is that in McKenzie's three kinds of case, the connection between economics and the economy is supposed to be implemented by the *"use"* of economics by economic actors. But using an economic model goes well beyond just recognizing it uttered or written down properly and understanding its meaning in the context. *Use involves taking further action.* Many kinds of further activities are needed, such as informing, learning, applying, arguing, implementing, predicting, calculating, estimating, negotiating, persuading, mobilizing resources, investing, agreeing, solving problems, winning conflicts – by a variety of academic and non-academic agents in the course of time. This undermines the idea that saying so non-causally makes it so.

It is no news that economic theorizing can have, and actually does have, many kinds of consequences for the economy. But these consequences largely flow through indirect causal rather than direct constitutive connections. The popular phrase used is that the economy is "shaped" by economics. Literally speaking, economic theories do not shape the economy. Nor does economic inquiry. People do. In their various roles (as policymakers, students, investors, entrepreneurs, workers, consumers) people are exposed to the results of economic inquiry and they learn, directly or indirectly, about the contents of economic theories, explanations and predictions, and are inspired by them, perhaps by being persuaded by the proponents, so as to modify their beliefs and perhaps their motives. These modified beliefs and motives make a difference to their behaviour, and this has consequences for the economy. The flow of these complex connections is a matter of indirect causal influence rather than direct constitution (see Mäki 2002).

The same holds also for MacKenzie's strongest form of "Barnesian performativity" whereby the use of a model makes it more true, makes it more closely correspond to the world. His example is the famous Black-Scholes-Merton model and the formula of option pricing derived from it. The formula has indeed been very important in informing and guiding practices in options markets. These practices "in their turn helped to create patterns of prices of which the model was a good empirical description. In that sense, the performativity of the model was indeed Barnesian." (MacKenzie 2006b, p. 33) Again, there is no constitutive relationship here between the theoretical model and some empirical practices and patterns. If it happens that certain practices and arrangements and patterns in real world finance are in line with the Black–Scholes–Merton formula, this naturally does not mean that the theoretical formula or its uttering by those three and other academic scholars – or by practitioners in the world of finance – "performs" those practices, making them occur by constitution. They may occur because the theoretical formula has managed to travel from academic research to economic practice in the manner outlined above. The connections are causal.

Sometimes the role of economists is rather direct in contributing to the shaping of the economy. In such cases the economist acts like an engineer rather than a theoretician interested in explaining phenomena. This is so in the new "design economics" that is directed towards meeting the practical demand for designing well-functioning markets (for whatever, such as electricity or kidneys) while meeting some moral or other constraints (Roth 2002; see also Guala 2007 on this "builder" role of experimental economics).

4 Further Queries About Performativity: Austinian and Otherwise

Let me expand on the differences between the Austinian notion of performativity and that of the new sociological performativists. Consider the elements of '*X* performs *Y*' in the two cases. What performs what? In the new performativist case, economic theories (*X*) perform markets (*Y*) by being used, perhaps with effects. In the authentic Austinian case, speakers (*X*) perform actions (*Y*) in uttering performative sentences in suitable circumstances. In uttering "I promise to pay you, *Europe,* my debt" *Manasses* performs the action of promising her to pay his debt. Now how do the two kinds of case compare – theories performing markets and speakers performing actions?

Consider first the *Y* part: *what is performed*. It is essential for the functioning of financial markets that they involve numerous promises and agreements put in terms of Austinian performatives. This is performativity within markets. Market agents perform certain kinds of actions. Their speech acts constitute those actions, and those actions partly constitute markets (or constitute parts or aspects of markets). In particular, contracts are at the core of markets, and contracts are Austinian-performed. By suggesting this much, am I not implying that markets are performed? Yes, with a proviso. The proviso is that market agents' speech acts do not perform markets in toto; at most they perform just bits and pieces, or certain limited aspects, of markets. And these bits and pieces are far more limited than is suggested by the sociological performativity thesis.

Consider then the *X* part: *what performs*. The remarks above dealt with performativity within markets, with market agents performing actions. Much of the time, however, the new sociological performativists claim that economic theories (or economists) perform markets. And here we don't find even partial Austinian-performativity. Theories are said to "perform" markets by way of travelling through an institutional structure and ultimately being used by market designers and market agents in their practices. But theories are not utterings, and utterings of theories don't constitute markets, not even those aspects of markets that are Austinian-performed (those considered in the previous paragraph). Utterings of theories may have powerful consequences for practice, but these consequences are not constituted by those utterings.

Austin himself has a familiar distinction that highlights the difference I have stressed between constitution and causation. It is the distinction between illocutionary and perlocutionary speech acts. Uttering one and the same sentence can serve both purposes. In one case, a speaker performs an action; in the other, she brings about effects or consequences. In an illocutionary speech act, "In saying I would shoot him I was threatening him" while in a perlocutionary act, "By saying I would shoot him I alarmed him" (Austin 1962, p. 122). Threatening is not a distinct consequence or effect of "I'd shoot you" whereas alarming is. "I'd shoot you" *is* to threaten, and threatening may have the external effect of the target person becoming alarmed. The speaker performs an illocutionary act, and this performance may have perlocutionary effects that are separate from the performance. Those effects are not performed by the speaker. They come about through a causal process.

Francesco Guala suggests a way of defending what he considers "genuine" performativity, one according to which economics contributes to "the making of *homo economicus*" by shaping people's behaviour by virtue of its normative authority. Economic models typically postulate an image of economic agents in terms of behavioural powers and dispositions such as the rational pursuit of self-interest, and this provides normative guidance for actual behaviour. Guala believes this is akin to Austinian performativity: "Economics can shape behavior because it works in part as a *norm* for the agents in the market, just like the priest's utterance 'you are now man and wife' creates powers and obligations for the individuals involved in a wedding ceremony." (Guala 2007, pp. 152–153). But the first case is not quite like the second. In the second case, the priest's utterance constitutes the creation of a marriage with powers and obligations. In the first case, presenting ("uttering") an economic model with *homo economicus* in it does not constitute the creation of rational economic men in the actual world. It may have causal consequences for actual behaviour by way of inspiration and encouragement, suggesting principles and policies, possibly to be adopted by acting people. Economics may have normative authority, but no Austinian-performative illocutionary force.

There is yet another important difference. One of the core connotations of 'perform' is that when performing an action or a task (or perhaps a play), one accomplishes it through and through, completely from the beginning to the end. In virtue of the constitutive power involved, the Austinian notion of performative speech act has this connotation, while even the strongest "Barnesian performativity" does not have it. It is enough for the latter that the world becomes *more similar* to the model, so that the model becomes *more true*. By contrast, if I say "I promise to pay you 100 euros tomorrow" I will have thereby given a promise to pay you exactly 100 euros tomorrow, not 75 euros nor any time later than tomorrow nor anything less than a full promise. Such compromises would be analogical to "more similar" and "more true" in a model becoming less than completely true in consequence of being used. This is yet another reason why Barnesian performativity fails to be an instance of genuine performativity.

5 Performativity in the Reverse Direction?

I have suggested that it is hard to see the sense of talking about the performativity of economics in the way the sociological performativists mostly do. But there is another way of speaking that might make a little more sense. Instead of saying that economic theory performs the economy, we could reverse the direction by saying that *the economy performs economic theory*. So in our formula "*X* performs *Y*" *X* and *Y* would swap places. However, not even this would give us Austinian performativity; *X* would not perform action *Y* by uttering some performative. Instead, the relevant exemplars of this idea include performing a play in theatre and performing a musical composition, as in "Dramaten performs Strindberg's *Creditor*" and "Helsinki Philharmonic orchestra performs Sibelius's *Kullervo*".

On this suggestion the analogy is between what could be generally considered scripts: written plays and musical compositions (such as *Creditor* and *Kullervo*) on the one hand, and economic theories and models (or what is derived from them, such as the Black–Scholes–Merton formula for option pricing) on the other. These scripts are then performed in concert halls and markets, respectively. The connotations are obvious: to perform a script is to execute, to carry out, to implement.

While this would make sense, or at least more sense, I am not proposing that this way of speaking should be adopted when studying the ways in which economics relates to the economy. It is hard to see any well-grounded motivation for it, just as it is hard to see any such motivation for the use of 'performativity' in ways that I am criticizing here.

6 So What's the Problem?

Someone might argue that one is free to use language as one wishes, so if Callon and MacKenzie want to use 'performativity' in a new way for their purposes, there should be no complaint. Yet I do complain. I think the use of 'performativity' for characterizing the economics/economy relationship is not only unhelpful but harmful for several reasons. It obscures the pretty well established authentic Austinian meaning of the term and, with no motivation given, replaces it with a number of other poorly defined meanings. Fashionable and weakly controlled terminology here contributes to conceptual clutter. More specifically, it obscures the important difference between causal and constitutive relations, thereby reducing the conceptual capacities of the suggested framework in identifying different sorts of detail in the ways in which economics and the economy are related.

As we have seen, the complex process through which these influences travel from economics to the economy may, and typically does, contain Austinian-performative links or moments. These may include opening and adjourning meetings, setting new rules and laws, drawing contracts and quitting from them, bidding and requesting, endorsing and questioning proposals, announcing intentions and decisions, expressing flattery and denouncement, warning about risks and congratulating on successful strategies, and so on. In other words, there are constitutive links in the overall causal chain that connects economics and the economy. These constitutive links deserve to be recognized as being such, and for this to be possible, distinct concepts are needed.

Consider one of MacKenzie's passages: "Option theory was thus used as a guide to trading and to hedging, and also to legitimate option markets. For these uses to qualify as effective performativity, economic processes with the theory being used must differ from processes without it being used." (2006a, p. 39) This suggests that the overall relationship – between option theory and, say, patterns of option prices – is causal (and analyzable in terms of counterfactuals and difference making). But the broad causal connection includes elements such as trading and hedging, and these involve Austinian-performative components such as contracting and promising and rule-setting.

Compare this to what might appear to be a somewhat different kind of case: physics and its engineering applications. Given their definitions of the term (e.g. in terms of MacKenzie's formulation of his fourfold typology), shouldn't the new sociological performativists be prepared to say that *physics performs bridges and computers*? Physics is being used, and it makes a difference. Using physics for designing and building bridges and computers is a causal process. Yet, it too contains Austinian-performative elements simply because the process is also a social one in that it involves things such as meetings, contracts, and promises required for organizing and resourcing the social and physical process of production. But it would make little sense to say that these elements are performed by physical theories; they are performed by people (many of them not physicists) acting within institutional structures, collectively contributing to the production of bridges and computers with the help of physics.

The damage done by the new sociological performativists is that by putting the vast mixture of ingredients and aspects of the theory-world relationship in one big box called 'performativity' they lose the capacity of recognizing genuine (Austinian) performative elements from other sorts of relationships. Their impoverished framework misses a conceptual resource that could be fruitfully used for a nuanced analysis of the complexities and diverse aspects in the economics-economy relationship. This is nothing but scientific regress that calls for conceptual recuperation. Only some initial steps in this project have been taken above.

References

Austin, J.L. (1962). *How to do things with words*. Oxford: Clarendon.

Callon, M. (2007). What does it mean to say that economics is performative? In D. MacKenzie, F. Muniesa, & L. Siu (Eds.), *Do economists make markets? On the performativity of economics* (pp. 310–357). Princeton: Princeton University Press.

Didier, D. (2007). Do statistics 'perform' the economy? In D. MacKenzie, F. Muniesa, & L. Siu (Eds.), *Do economists make markets? On the performativity of economics* (pp. 276–310). Princeton: Princeton University Press.

Guala, F. (2007). How to do things with experimental economics. In D. MacKenzie, F. Muniesa, & L. Siu (Eds.), *Do economists make markets? On the performativity of economics* (pp. 128–162). Princeton: Princeton University Press.

MacKenzie, D. (2004). The big, bad wolf and the rational market: Portfolio insurance, the 1987 crash and the performativity of economics. *Economy and Society, 33*, 303–334.

MacKenzie, D. (2006a). Is economics performative? Option theory and the construction of derivatives markets. *Journal of the History of Economic Thought, 28*, 29–55.

MacKenzie, D. (2006b). *An engine, not a camera. How financial models shape markets*. Cambridge: MIT Press.

MacKenzie, D. (2007). Is economics performative? Option theory and the construction of derivatives markets. In D. MacKenzie, F. Muniesa, & L. Siu (Eds.), *Do economists make markets? On the performativity of economics* (pp. 54–86). Princeton: Princeton University Press.

MacKenzie, D., & Millo, Y. (2003). Constructing a market, performing theory: The historical sociology of a financial derivatives exchange. *American Journal of Sociology, 109*, 107–145.

MacKenzie, D., Muniesa, F., & Siu, L. (Eds.). (2007). *Do economists make markets? On the performativity of economics*. Princeton: Princeton University Press.

Mäki, U. (2002). Some non-reasons for non-realism about economics. In U. Mäki (Ed.), *Fact and fiction in economics. Realism, models, and social construction* (pp. 90–104). Cambridge: Cambridge University Press.

Mäki, U. (2008). Scientific realism and ontology. In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed., Vol. 7, pp. 334–341). Basingstoke: Palgrave Macmillan.

Mäki, U. (2012). Realism and antirealism about economics. In U. Mäki (Ed.), *Handbook of the philosophy of economics* (pp. 3–24). San Diego: Elsevier.

Roth, A. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica, 70*, 1341–1378.

Beyond Motivation and Metaphor: 'Scientific Passions' and Anthropomorphism

Lisa M. Osbeck and Nancy J. Nersessian

Abstract We align with other challenges to the idea that emotion-free science, even in principle, is a productive scientific value. We emphasize that emotion can be seen to have important functional benefits for the research scientist and the wider science. Here we analyze the function of anthropomorphic expressions from practicing bioengineering scientists and claim that anthropomorphisms can be an indirect or roundabout indicator of emotional experience. We claim that the attribution of emotional states through anthropomorphism contributes to the motivation, interest, and attention of the researcher and may carry implications of agency, such that objects central to problem solving are imbued with agency and transformed into working partners with the research scientist in cognitive practices toward shared and individual problem solving goals.

In this paper we take up the relation of emotion to scientific reasoning and align with other challenges to the idea that a cold and passion-free science, even if possible or even in principle, is a productive scientific value. We claim, instead, that emotion, one form of which is expressed through anthropomorphism, has important functional benefits for the research scientist and for the wider science of which the scientist's work is a part.

The functional benefits of anthropomorphism are of two related kinds: First, the attribution of emotional states through anthropomorphism reflects implicit emotional processes that contribute to the motivation, interest, and attention of the researcher in relation to the objects and entities central to the laboratory's research

L.M. Osbeck (✉)
Department of Psychology, University of West Georgia, Carrollton, GA 30118, USA
e-mail: losbeck@westga.edu

N.J. Nersessian
School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30308, USA

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_37,

projects. Second, the attribution of emotion carries attributions of agency. That is, objects central to the practice of the scientist are imbued with agency (functionally so) through anthropomorphism, such that the research scientist comes to view them as transformed into partners in the cognitive practices aimed at achieving problem solving goals. We call this relationship between researcher and artifact "cognitive partnering."[1] We base the notion of cognitive partnering and our related analysis of anthropomorphism on our ten-year investigation of bio-engineering sciences laboratories and (most recently) integrative systems biology. In this paper our claim is that emotional expressions, including anthropomorphisms, may be analyzed in terms of their functional role within and as reflective of the normative structure of the distributed cognitive-cognitive system – the laboratory - laboratory in which they occur.

1 Science as "Cold"

The idea that emotions impede rational powers is typically assumed to originate with the chariot allegory in Plato's *Phaedrus* (Solomon 1993; Thagard 2008), depicting inevitable conflict created by the demands of opposing agencies. That this reading of reason and emotion as juxtaposed or in conflict represents a misreading of the very allegory to which it is credited is a point worth making; the more salient point is that the separation was problematically reified in the nineteenth century in particular (Dror 2009).

Emotion is viewed as an impediment to science when it introduces a set of particular or "individuating" processes belonging to the scientist that can introduce bias. Of late some qualifications have been made. Daston and Gallison (2007) suggest that passion for scientific research in general is tolerated and viewed as compatible with objectivity but that "passionate preferences for one's own theories and speculations" are not (p. 380). White acknowledges the "familiar tropes such as the pleasure of knowledge, the passion for truth, the thrill of discovery" (2009, p. 792), but notes that emotion is merely tolerated as a necessary byproduct, not as integral to science itself. Earlier, Polanyi (1973/1958) raised a more direct challenge to the traditional view by arguing that a view of science as detached or neutrality as necessary to the authoritative grounding of science is misleading and destructive: science without passion would be science without directed interest. In the wake of the more recent "affective revolution" in cognitive science (Haidt 2007,

[1] Our notion of cognitive partnering might appear to be the same as that of the "actor network" introduced by Bruno Latour (1987). There are significant differences between 'partners' and 'actors'. First, unlike Latour's actors, not all partners are equal. Human partners (agents) ascribe agency to salient artifacts, but true agency (on our view) requires intentionality, and so artifacts can perform cognitive functions in the system and exhibit independent behaviors, but are not themselves agents.

p. 998), a heightened appreciation for the profound impact of emotion on cognitive tasks and processes is evident (Bechara 2004; Damasio 1994, 1999; McAllister 2005; Thagard 2008).

Although these efforts have not translated into a widescale revisioning of scientific cognition, there is evidence of considerable rethinking of the place of emotion in science in both cognitive science and science studies. A Focus section of Isis on "*The emotional economy of science*" for which White's paper provides the introduction is exemplary of attention to emotion in the history of science community. In turn, White notes that some of the historical work on emotion in science draws from new models in cognitive science, neuroscience, and sociology "which allow for conjunctions between the psychological, the social, and the material, as well as between the emotional and the rational," some offering "critical revisions" of these relations (p. 793). In truth, however, it is difficult to identify models that allow full integration of all of these dimensions (psychological, social, material, emotional and rational). Rather, as White points out, models positioning reason and emotion as integral to one another are those that identify emotions as bodily processes (Damasio 2003; LeDoux 1998; Lakoff and Johnson 1999), with emotion, like reason, "shaped by the body" (Lakoff and Johnson 1999, p. 5). When the emphasis is on cognition at the organizational or system level, emotion is less clearly part of the mix. Thus there is little discussion of emotion in the context of distributed cognition, for example. The study of emotion as culturally or historically produced (Parrott and Harré 2001) similarly leaves unclear how we might theorize the embodied emotional processes of the particular scientist. There is a need for a coherent account of distributed "emotional cognition" in science practice.

2 Discovery Versus Everyday Practices

One place where even personal passion *has* been allowed in scientific discourse is in relation to the process of discovery, in the "overwhelming elation felt by scientists at the moment of discovery" (Polanyi 1973/1958, p. 134). Thagard has written recently on the neural processes underlying the "the wonderful AHA! experiences that creative people sometimes enjoy," (Thagard and Stewart 2011, p. 1), calling it a "treasured" experience for scientists (p. 2). But as Polanyi acknowledges, the passions of *discovery* generally are not assumed to interfere with the objectivity of scientific process. They are most frequently assumed to be "mere psychological by-play" of scientific reasoning. Their function, if one is to be admitted, is merely motivational in nature, sustaining the scientist through laborious procedures and inevitably frequent failures with the lure of occasional elation.

Far less is understood about the nature and function of emotion in the "ordinary, artful accomplishments" (Garfinkel 1967, p. 14) of everyday scientific practices, many of which are removed from the thrill of discovery. Polanyi recognized that emotions are not only an inescapable feature of the everyday activities of science, including the very act of observation, but that they are essential to its successful operation. He describes different categories of passion beyond discovery, but his method does not allow an exploration of concrete instances of practice as lived. Even if emotion, like reason, is an embodied achievement, how it is produced, displayed, and how it functions in relation to the wider goals of the epistemic community remains in question. To address this question we require the kind of analysis made possible by ethnography and ethnomethodology, which also enables the consideration of the function of emotion in science at the level of the system, the "distributed problem-solving space" (Nersessian et al. 2003) such as the working, evolving research laboratory.

3 The Resistance Problem

A problem confronting any effort to understand the role of emotion in day to day science is that scientists themselves resist the very idea that emotion could be mingled with their practices. To acknowledge the very presence of emotion in scientific reasoning is to leave oneself open to the charge that one's reasoning is *led* by emotion.

The implications that follow from scientists' resistance are largely methodological, namely that it is important to identify the presence or function of emotion in scientific practice in somewhat 'covert' ways. We have argued that one roundabout solution is to pay attention not only to overt expressions of emotion but to look at metaphorical and figurative expressions in scientists' descriptions and anthropomorphisms involving an attribution of emotional states to objects, artifacts, and devices (Osbeck and Nersessian 2011). Here we examine examples of anthropomorphism in more detail to consider the function these kinds of expression might serve in the overall problem solving configuration of the research laboratory.

4 Context and Method

Our analysis takes place within a multi-year ethnographic investigation of cognitive and learning practices in four innovation focused bioengineering sciences laboratories. We draw from three labs here: tissue engineering, neural engineering and integrative systems biology (added recently).

4.1 Lab A

Lab A aims to understand mechanical dimensions of vascular cell biology and to engineer living substitute blood vessels for implantation in the human cardiovascular system. Intermediate research problems included, for example, designing and building living tissue – "constructs" – that mimic properties of natural vessels; creating endothelial cells (highly immune-sensitive) from adult stem cells and progenitor cells; and designing and building environments for mechanically conditioning constructs.

4.2 Lab D

Lab D's research problems are to understand the mechanisms through which learning occurs in networks of living neurons, potentially to use this knowledge to aid neurological deficits. Daily working problems included developing ways to stimulate, control, record, and image cultured neuron arrays (locally called "the dish") and designing and constructing feedback environments (robotic and simulated) through which the "dish" could learn.

4.3 Lab C

Lab C, a systems biology lab, conducts both computational modeling and wet lab experimental research. It seeks to understand cell signaling dynamics in Redox (reduction – oxidation) regulation of the cells in dynamic immunological contexts (e.g. cell aging, cancer). Diverse research projects are pursued in daily work, such as developing a systems-based and quantitative model of chemotherapeutic drug resistance in acute lymphoblastic leukemia cells.

Collectively, our research group has conducted numerous hours of *in situ* field observations of the researchers at work and of lab meetings, and hundreds of unstructured interviews. Our framing assumption in approaching data collection and analysis is that the cognitive practices of the laboratory are both situated in problem contexts and distributed across systems of interacting persons, artifacts, instruments, and traditions. By situated we refer to a view of learning and problem solving as enabled or constrained by the particular features of the environment in which they occur, including the social environment. By distributed, we mean that we regard brain and environment as co-constituting a single complex system of interacting processes. In so framing our study of the laboratories as situated and distributed, we connect it to other investigations of real-world problem solving that implicate the environment in cognition in important ways (Greeno 1998; Hutchins 1995; Resnick et al. 1997). Consistent with this tradition, we use interpretive methods to provide a thick description of the complexities of science practice as it takes place in its natural settings. The use of individual interviews with researchers at different levels of expertise and from different disciplinary backgrounds enables us to analyze how the particular learning history and affective style of the researcher might contribute to the overall situation in which the research is conducted.

4.4 Coding

Broadly consistent with the aims of grounded theory, we have been coding interviews with respect to our research questions, so as to enable core categories to emerge from interview data and remain grounded in it (Glaser and Strauss 1967; Strauss and Corbin 1998). The focus of our analysis has been cognitive practices; however, in the process of coding we found there to be numerous expressions of emotion that appeared related to these practices.

5 Anthropomorphizing Expressions

5.1 Attributions of Emotional States: Cells and Networks

A form of emotional expression we found especially interesting and to which we direct our focus here is the attribution of emotional states to the entities and artifacts central to the problem solving practices in the laboratories: anthropomorphisms. By anthropomorphism, of course, we refer to the attribution of human qualities, states, or characteristics to non-human things, particularly as an explanation for the behavior or potential experience of that thing. The most striking and theoretically important example of this practice is the attribution of happiness to cells across all the labs that use them in research. For example:

> A22: *You would want to have the smooth muscle cells on the inside so that you can – so that the endothelial cells will be happy.*
>
> D4: *Cell density is important, because cells survive more if they, if they're connected to each other. A lone cell by itself is not very happy.*
>
> C10: "And so sometimes the cells would get lysed 'cause they would see, on the corners the cells like see more force,... so that they weren't happy.

Researchers across levels of expertise exhibit this practice of attributing happiness to cells, from the director to the undergraduates.

Especially noteworthy is an interview with D4 which reveals a normative aspect of the concern with happy cells (neurons), an expectation for researchers to keep the cells happy, and to *care* about keeping them happy. The context is an early interview with a graduate student; it reflects the ethnographer's effort to understand the lab's basic research practices:

> I: You guys talk about the neurons with the undergrads all the time... so, why do you guys care so much about neurons?
>
> D4: *If you don't essentially care about the nodes in the network, ...you gotta care about them!*
>
> *...So they make up the network, each of them has a part to play, in the network property, so you want to keep as many as you can. You know, because they make up essentially, as I said, they make some basic rules for the way the network works, so you want to keep them happy.*

The implication here is that it is not only the happiness of the cells as a collective that is important but even the happiness of individual cells, given that each of them contributes in a particular way to the functioning of the whole. In a later interview with another interviewer, D4 helps to clarify that "happy" cells are actively forming connections – the patterns of which are the essential focus of all research activity in the neuroengineering lab:

> D4: *. . . this flow of activity is basically because of the connectivity of the network . . . . But then there are some cells which are always active like this guy here, and that guy there. They are just happy firing all over on their own . . . . maybe they're two cells connected to each other, they go bing-bing-bing all the time, I don't know.*

6 Functional Significance of Anthropomorphism

The relevant question what might be the functional significance of anthropomorphic attributions within the context or situation of the laboratory and its goals. One means of analyzing the function of anthropomorphisms is to consider their functional value in other contexts of human activity.

The attribution of emotional states in the case of religious belief functions as a form of demand, a demand for action adhering to the moral standards of the community, as was first analyzed by Spencer (1858). In the laboratories, attributing emotional states to cells similarly functions to demand care and attention. Cells could die if not properly cared for. Dead cells cannot be cultivated and adapted to research goals. If cared for, if kept 'happy', cells will thrive and form connections. As in the context of religion and the demand for adherence to the moral order, anthropomorphism is a potent, effective way to insure that researchers care for their cells, and in so doing advance their own research goals and those of the laboratory community.

The demand for care is similar in this regard to a second human context in which anthropomorphism prevails, that of pet ownership. Serpell (2006) addresses the evolutionary significance of anthropomorphism in pet ownership, seeing it as "a powerful transformative force that has not only molded the behavior and morphology of our animal companions in unprecedented ways but also, through them, enhanced our own health and well-being" (p. 122). One function of anthropomorphism in pet ownership is to secure care, ensuring moral responsibility as discussed in relation to religious anthropomorphism. But a further step in the case of pets is that reciprocity of well-being is established through anthropomorphism; there are clear benefits for the pet owner and perhaps only equivocal benefits for the pet.

The few remarks Polanyi makes about anthropomorphism in the context of discussing scientific passions support the functional value of anthropomorphism. Anthropomorphism imbues interest and meaning to the array of potential facts in the researcher's domain: "Living animals are more interesting than their dead bodies; a dog is more interesting than a fly; a man more interesting than a dog" (Polanyi 1973/1958, p. 138). In general, "charging" objects with emotion functions to "affirm that something is precious"... "more particularly, that it is precious to science" (p. 134). Polanyi understands emotions as having a normative component in the practice of science: they may be judged as right or wrong as appropriate to the inquiry. Certain emotions are "rightful" in scientific reasoning (p. 134). This provides one means of understanding the norm of "caring about cells" in laboratories. Attributing emotion to cells, which elicits care, makes them more interesting to the researcher, imbues them with compelling meaning. Greater interest secures greater focus, enhances the potential for innovative thinking. Attributing emotional states to cells, then, not only offers benefits to the researcher but to the science of which the researcher's particular problem solving goals are a part. Further, meaning may be established by consulting one's own experience in attributing emotional potentiality to objects. There is a "feeling with" cells in this case that establishes the grounds for "working with" them in relation to the research questions.

6.1 Implications of Agency

Attribution of human qualities to objects or animals, particularly as explanation for behavior, relates to what Dennett has called "intentional stance:" interpreting behavior as if the entity were an agent freely choosing, believing, and desiring (Dennett 1996; Duffy 2003). We found this intentional stance in relation to cells, the network, and other artifacts:

> D6: *Cells make a lot of decisions with whom they want to connect with.*
>
> A11: *Uh, well, the cells once they are in the constructs will reorganize it and secrete a new matrix and kind of remodel the matrix into what they think is most appropriate.*
>
> D4: *Well, that way I'm strengthening that particular pathway, so the network would prefer to always excite these two cells in a certain way after my modifying input.*
>
> D6: *So this computer is listening, and what it's listening to you can think of as the motor output.* A11: *It [bioreactor circuit board] sees voltages in different ways.*

The attribution of agency or intention to cells and other artifacts of the practice accompanies what we have called perspective taking, the ability to take the perspective of the cell, for example, in perceptual or other experience:

> A10: *But from a cell=s perspective, the cell sees basically a flat surface. So to the cell - it has no idea that there's actually a curve to it. The cell, when it looks around, just sees a flat surface. Just like we think the earth is flat.*

7 Cognitive Partnering

The attribution of emotional states, assignment of agency, and perspective taking enable and reflect what we have called *cognitive partnering*, cooperative participation within an epistemic culture that enables or sustains particular problem-solving strategies. Although cooperative participation also takes place with other researchers, we are most intrigued by partnering as a kind of sympathetic engagement with artifacts and devices, a tacit sense of working together with them toward problem-solving goals.

The sense of working together, in this case by modifying input, is expressed in the latter half of the passage earlier referenced about the network's "preference:"

> D4: *I'm strengthening that particular pathway, so the network would prefer to always excite these two cells in a certain way after my modifying input. Maybe they weren't before my modifying input, and after my modifying input the pathway becomes stronger.*

Although there are instances in which it is extended to lab built artifacts, notably simulation devices and instruments, cognitive partnering in these labs is most prevalent in relation to cells. Of central importance is the dynamic tension or *resistance* deriving from their status as living objects and the fact that they are central to the work of each laboratory. They do not always behave as researchers expect or want them to do, and they can die. If not kept happy, they interfere with the researcher's cognitive goals.

> D4: *Pfft, you keep them happy by feeding them, by taking care of them, hopefully stimulating them* [in a motherly condescending voice-note from transcriber] *and telling them to do something! I don't know what to do to make them happy. I don't know how make them happy, that'll make my neurons happy* [points to head]
>
> I: Hah, to make your neurons happy, your brain neurons happy.
>
> D4: *So, my experiment is very.. so we're writing a paper about, about this burst suppression* [something thought necessary to make the cells happy].

Here, the researcher directly ties the "happiness" of her own neurons, which we interpret as meaning her own scientific thinking and problem solving, to her frustrated attempts to find a way to keep the cells happy. Importantly, cognitive partnering with cells constitutes a *transformation in relational stance* such that the researcher cares more about their welfare and takes on more responsibility for their well-being.

I2 Well I know you, you often, you often refer to your cells as "my cells", you seem like you've gotten some sort of... you like them!

> A8 *Relationship? They're my children!*
>
> I2 Do you ever think of them that way?
>
> A8 *Oh yeah, when I was first being trained, the woman who trained me, everyone gets trained on cell culture by someone, right so.... She called 'em children. I think that's a very good analogy because you have to feed 'em, you have to keep 'em alive, you have to take care of them, you know, and they, they eat, and they get hungry, and... I do call them mine, because... I think of them that way.*
>
> I2 It hard to think of a, a ... new piece of rubber like that... or something like that, right?
>
> A8 *Well then you think of it as property, but not as much as something you're taking care of.*

In the following passage from Lab C, the implication is not only that cells must be cared for, but that researchers themselves pose a threat to the well-being of cells and that they must anticipate and manage this threat. The attribution of happiness also enters in to the concern with the cells' protection and survival:

> C11 *When you're doing cell work–you have to think of the fact that you're not only trying to protect yourself, you're trying to protect your samples from you. You. Like you are the biggest risk of infecting your cells and completely messing them up– You have to keep them happy. In so much as single cells can be happy. You protect yourself and you keep them happy. And then after that you also have to remember all the time to be safe.*

The researcher here is a source of intrusion and potential threat to the cells and her own safety is secondary to their well-being and happiness. The transformation that constitutes cognitive partnering thus has both cognitive and moral implications.

8 Conclusions

In all of these contexts and examples, anthropomorphism serves important functions. Our analysis of the functional significance of anthropomorphic expressions in biomedical engineering laboratories is in keeping with what we have identified as locally normative component to the concern with happy cells (neurons) in the laboratories: an expectation for researchers to keep cells happy and to care about keeping them happy. The interpretation we offer differs from that of Knorr-Cetina, who analyzes the function of anthropomorphism to be principally metaphorical, aiding communication and understanding, even though she acknowledges anthropomorphism to express a "reconfigured order" (of relations, e.g. subject and object) in a context of science practice (1999, p. 284).

Questions might be raised concerning the cultural generalizability of our claims, whether "happy cells" might be an artifact of the peculiarities of American culture. It bears stressing that laboratory science is increasingly an international affair. Participants in our studies included graduate students from India, Taiwan, Sweden, France, and Nigeria. Whether anthropomorphic expressions would be used with similar frequency and would take on a normative dimension in laboratories situated outside of the US is a question to be settled with additional empirical study.

Returning to what we identified earlier as a problem of resistance confronting the researcher interested in emotional processes in science laboratories, some researchers from laboratories in our new study became aware of and read a paper on the biomedical engineering labs that focused on the forms of emotional expression and their implications, including some of the frequent references to happy cells and other anthropomorphisms considered here. The following exchange takes place in the course of a conversation about the necessity of getting a machine serviced. The manager imparts emotion and agency to the machine, then reacts to the fact that she has done so:

> I How often do you get the machines serviced?

> C11 *That actually depends on how often we run it. If we don't run it, it will break down easily, or it breaks down more easily.*

I. So you need to run it?

C11 *Yeah, it likes to be run ... It makes it happy.*

I Machines also need to be kept happy?

C11 *I should stop doing that!*

I Wait, you should stop doing what?

C11 *What's the word for it? When you, anthropomorphize?*

I Ah, Right, right.

C11 *If it's too life-like, it's creepy. But I keep doing it with the cells and the machines.*

It is difficult to determine whether this researcher's self-monitoring is better understood as resistance to the contaminating effects of emotion or to the prevalent perception of anthropomorphism as a form of sloppy linguistic practice. Daston and Mitman's observation that anthropomorphism is usually "applied as a term of reproach, both intellectual and moral" is useful to consider. Here the history of the word, its application in other contexts of human practice is relevant and helps to explain contemporary attitudes toward it: "Originally, the word referred to the attribution of human forms to gods, forbidden by several religions as blasphemous. Something of the religious taboo still clings to secular, modern instances of anthropomorphism, even if it is animals rather than divinities that are being humanized (2006, p. 2).

Despite the resistance in and out of the scientific community, our point is attribution of emotional states and assignment of agency enables cognitive partnering, a stance of engaged participation with the objects and artifacts of practices that enhances all dimensions of scientific problem solving.

Acknowledgement This research was supported by the US National Science Foundation grants REC0106733, DRL0411825, DRL097394084.

References

Bechara, A. (2004). The role of emotion in decision making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition, 55*(1), 30–40.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Grosset/Putnam.

Damasio, A. (2003). *Looking for Spinoza: Joy, sorrow, and the feeling brain*. New York: Harcourt.

Daston, L., & Gallison, P. (2007). *Objectivity*. Cambridge: MIT Press.

Dror, I. (2009). How can Francis Bacon help forensic science? The four idols of human biases. *Jurimetrics: The Journal of Law, Science, and Technology, 50*, 93–110.

Dennett, D. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books.

Duffy, B. (2003). Anthropomorphism and the social robot. Special issue on socially interactive robots. *Robotics and Autonomous Systems, 42*(3–4), 177–190.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs: Prentice-Hall.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Greeno, J. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*, 5–26.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 18*(316), 998–1002.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge: MIT Press.

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge: Harvard University Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh*. New York: Basic Books.

Latour, B. (1987). *Science in action*. Cambridge: Harvard University Press.

LeDoux, J. (1998). Fear and the brain: Where we have been, and where are we going? *Biological Psychiatry, 44*(12), 1229–1238.

McAllister, J. (2005). Emotion, rationality, and decision making in science. In Petr, H., Valdés-Villanueva, L., & Westerståhl, D. (Eds.), *Logic, methodology and philosophy of science: Proceedings of the twelfth international congress* (pp 559–576). London: King's College.

Nersessian, N., Kurz-Milcke, E., Newstetter, W., Davies, J. (2003). Research laboratories as evolving distributed cognitive systems. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pp. 857–862.

Osbeck, L., & Nersessian, N. (2011). Affective problem solving: Emotion in research practice. *Mind and Society, 10*, 57–78.

Parrott, W., & Harré, R. (2001). Emotions in cultural contexts in space and in time [Special issue]. *International Journal of Group Tensions, 30*(1).

Polanyi, M. (1973/1958). *Personal knowledge: Towards a post-critical philosophy*. Chicago: University of Chicago Press.

Resnick, L., Pontecorvo, C., & Säljö, R. (1997). Discourse, tools, and reasoning. In L. Resnick, R. Säljö, C. Pontecorvo, & B. Burge (Eds.), *Discourse, tools, and reasoning: Essays on situated cognition*. Berlin: Springer.

Serpell, J. (2006). People in disguise: Anthropomorphism and the human-pet relationship. In L. Daston & G. Mitman (Eds.), *Thinking with animals* (pp. 124–136). New York: Columbia University Press.

Solomon, R. (1993). The philosophy of emotions. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (pp. 3–15). New York: Guilford.

Spencer, H. (1858). The use of anthropomorphism. In *Essays: Scientific, political, and speculative* (pp. 430–435). London: Woodfall and Kinder.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research techniques and procedures for developing grounded theory* (2nd ed.). London: Sage.

Thagard, P. (2008). *Hot thought*. Cambridge: MIT Press.

Thagard, P., & Stewart, R. (2011). The AHA! experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35*, 1–33.

White, P. (2009). Introduction. Focus: The emotional economy of science. *Isis, 100*(4), 792–797.

A Way of Saving Normative Epistemology? Scientific Knowledge Without Standpoint Theories

Maria Cristina Amoretti and Nicla Vassallo

Abstract Feminist standpoint epistemologies of the sciences must be acknowledged to have significant merits. However, as we have already argued, the very notion of standpoint – being intrinsically linked to the notions of better epistemic reliability, privilege, or advantage – brings with it an unavoidable dilemma: it forces its defenders to choose between essentialism (or at least its negative and dangerous consequences) and regarding all standpoints as equal. Moreover, we have also noted that there is no reason to appeal to any feminist standpoint epistemology of the sciences to retain all of its more significant merits. Given the importance of the debate, this paper aims to rebut possible objections that standpoint theorists may advance against the general argument from essentialism that we defend and to show that there is no effective way of supporting a genuine (i.e., normative) standpoint epistemology.

1 Standpoint Theories and the Challenge of Essentialism

The notion of standpoint is of great importance among feminist epistemologies and epistemologies of the sciences, but it is neither clear nor univocal because there are many different ways of grounding and developing any specific standpoint theory (Harding 2004). For this reason, it is better to talk about standpoint *theories*, in the plural. Even granting the heterogeneity of standpoint theories, however, it remains

M.C. Amoretti (✉) • N. Vassallo
Faculty of Humanities, University of Genoa, Via Balbi 4, 16126 Genoa, Italy
e-mail: cristina.amoretti@unige.it; nicla.vassallo@unige.it

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_38,

possible to find some pivotal features in common among them.[1] Janet Kourany recognises at least four:

> Certainly, [standpoint theory] claims [1] that all knowledge is situated, positioned in a particular time and place; [2] that where power is organized hierarchically – for example, by class or race or gender – persons can achieve only partial views of reality from the perspective of their own positions in the social hierarchy; [3] that the view from the perspective of the less powerful is far less partial and distorted than the view from the perspective of the more powerful; [4] that that superior view or 'standpoint', however, must be discovered and developed through a collective process of political struggle and consciousness-raising. (Kourany 2009, p. 210)

The first and the third features are particularly relevant for our argument. On the one hand, standpoint theories correctly place their emphasis on the fact that knowledge in general and scientific knowledge, in particular, are situated: our understanding of the natural and social worlds (partially) depends on our specific perspective on that world. Importantly, this perspective is not automatically assimilated but is the result of critical analysis and must be engaged, struggled for, and in this way, achieved. On the other hand, according to standpoint epistemologists, knowledge of marginalised and/or subordinated social groups is considered less partial, less distorted, and less false than knowledge of dominant groups. In other words, marginalised and/or subordinated social groups are estimated as more epistemically reliable, privileged, or advantaged than dominant groups. Women in a patriarchal society constitute a marginalised and/or subordinated group, and thus, according to feminist standpoint theorists, their social location, once rightly appreciated, makes women's knowledge more epistemically reliable, privileged, or advantaged than knowledge of other groups. This reasoning, which we have applied to knowledge in general, may work also for scientific knowledge because, for the most part, scientific fields and endeavours are dominated by men and are thus rich in (implicitly or explicitly held) androcentric biases.

Elsewhere, we have argued that the very notion of standpoint – being intrinsically linked to the notions of better epistemic reliability, privilege, or advantage; that is, to the idea that certain (marginalised and/or subordinated) groups and their corresponding standpoints are more epistemically reliable, privileged, or advantaged than other (dominant) groups – brings with it an unavoidable dilemma for those who want to develop standpoint epistemologies and epistemologies of the sciences (Amoretti and Vassallo 2010a, 2011, 2012). Again, we consider the standpoint of women.

To establish and defend the thesis that the standpoint of women is more epistemically reliable, privileged, or advantaged than that of other groups, it is necessary to ground those notions in social and/or natural facts concerning women – such as women's cognitive style, women's common experiences, women's work

[1] For some key descriptions of the notion of standpoint, see (Bowell 2011; Brooks 2007; Haraway 1988; Harding 1991, 1993, 1995; Hartsock 1983; Potter 2007; Rose 1983; Wylie 2003).

conditions, women's practices, and so on[2] – and to explain why these facts, once appreciated, produce a better epistemic position. In this way, group membership (here, being a woman) would be defined by the very social and/or natural facts that ground the better epistemic reliability, privilege, or advantage of the group of women, facts that would inevitably point to the essence of "woman".[3] More precisely, there must be some causal connections between the social and/or natural position of a group and its better epistemic reliability, privilege, or advantage. Some characteristics (cognitive style, common experiences, work conditions, practices, and so on) of the social and/or natural position of the group of women make their standpoint more epistemically reliable, privileged, or advantaged. This means that these characteristics are necessary (even if not sufficient) to guarantee better epistemic reliability, privilege, or advantage to the group of women. Thus, these characteristics define membership for the group of women, at least as far as one wishes to consider women a more epistemically reliable, privileged, or advantaged group. Because essential properties determine group membership (Witt 2011), it is possible to conclude that these characteristics represent essential properties of the group of women (or, better, at least *some* of its essential properties because the possibility of other essential properties is not excluded).

As a regrettable consequence, such a move would lead to ignoring the evidence that each woman has her own particular identity, assuming the existence of a feminine "nature", and thus embracing essentialism (or at least – in the best scenario – its negative and dangerous consequences, such as conditions of exclusion or inferiority).[4] The obvious alternative solution is to reject the very notions of better epistemic reliability, privilege, or advantage; however, such a move would imply that

[2] It has been argued that some biological and cognitive differences occur between women and men that may be considered either as innate (Baron-Cohen 2003) or resulting from the different mechanisms of identity-formation faced by male and female children during the first period of their lives (Hartsock 1983). Whether innate or social, these (presumed) biological and cognitive differences surely become more evident in societies that are strongly gender-structured and that therefore impose different life experiences and social activities on women and men.

[3] When considering women, essentialism is the view that some properties are essential to all women, and any woman must necessarily have these properties to be a woman (Stone 2004). To put it another way, essentialism "says that members of the group of 'woman' have the same nature and thus there is a single universal womanness that all share" (Stoljar 2000, p. 177).

[4] Similar worries have been also advanced by Susan Hekman (1997) and Iddo Landau (2008). Moreover, as we have shown (Amoretti and Vassallo 2010a, 2011, 2012), such reasoning still applies to any fragmentation of the group of women in transversal social groups that intersect sex and gender with culture, race, social class, sexual preference, personal history, religion, and age (Collins 2000; Lugones and Spelman 1986). Admitting a plurality of standpoints is no less problematic; no matter how a transversal social group is defined, to explain and justify why the corresponding standpoint is more epistemically reliable, privileged, or advantaged than that of others, one must refer to social and/or natural facts concerning the transversal social group in question. Again, these facts would inevitably single out the essence of such a transversal social group.

all standpoints should be regarded as equal: in particular, the standpoint of women could not be considered more epistemically reliable, privileged, or advantaged than that of other groups.

Many efforts have been made to address this issue (Crasnow 2008; Harding 1997; Wylie 2003). However, as we will try to prove in the next section, none of these attempts is completely satisfactory, at least if the aim remains that of conceiving standpoint theories as genuine (i.e., normative) epistemologies.

2 The Argument from Essentialism: Some Objections

A general objection to the argument from essentialism is that it deliberately ignores, or at least underestimates, two basic tenets common to many, if not all standpoint theories: (i) those (e.g., men) who do not belong to the marginalised and/or subordinated group (e.g., women) may also adopt the standpoint of such a group; (ii) the mere fact of occupying a marginalised and/or subordinated social location does not automatically yield better epistemic reliability, privilege, or advantage because a standpoint is something that must be engaged, struggled for, and in this way, achieved (Crasnow 2008; Rolin 2009; Wylie 2003). We believe that these observations are correct, but nevertheless, unable to repudiate essentialism.

First, the claim that men can also adopt the standpoint of women is insufficient to reject essentialism. To achieve the standpoint of women, a man should at least recognise and appreciate the social and/or natural facts that confer better epistemic reliability, privilege, or advantage to such a group. The group "women" would continue to be defined by the social and/or natural facts on which its better epistemic reliability, privilege, or advantage is based. Thus, it would still be possible to single out the essence of "woman". A possible objection is that, according to feminist standpoint epistemologists, men can adopt the standpoint of women because there is nothing essential to the group of women. To respond to this objection, some clarifications are needed. Even if a standpoint is engaged, struggled for, and thus achieved, there must be something in its social and/or natural position that confers better epistemic reliability, privilege, or advantage independently of any training, engagement, and struggling (something that is necessary, common to all members of the group, and thus essential). If one would deny this claim, then training, engagement, and struggling would be sufficient in themselves, regardless of the social position occupied by a particular group. Hence, one would have no reason to distinguish between different social positions (Landau 2008). Moreover, if one gives up social positions, one must also renounce understanding the notion of standpoint in terms of situated knowledge (as many feminist standpoint theorists do: Harding 1991; Haraway 1988; Kourany 2009). To summarise the discussion of the claim that men can also adopt the standpoint of women, it is more accurate to maintain that men can "take advantage of" (rather than "achieve") the standpoint of women as far as they are able to recognise the characteristics (cognitive style, common experiences, work conditions, practices, and so on) that confer better epistemic reliability, privilege, or advantage to the group of women and that are common to all members of this group.

Second, even if one admits that a standpoint is something that must be engaged, struggled for, and in this way, achieved, the explicit admission remains that some social and/or natural facts concerning marginalisation and/or subordination are necessary – even if not sufficient – to obtain better epistemic reliability, privilege, or advantage (Crasnow 2008; Harding 1997). This move rules out the possibility that merely occupying a marginalised and/or subordinated social position would automatically yield better epistemic reliability, privilege, or advantage; however, it does not disprove essentialism. Generally speaking, an essential property of an object is a property that such an object must have, that it has necessarily, and that it could not lack and yet remain the object that it is (Mackie 2006; Robertson 2008). Thus, admitting that some social and/or natural facts are necessary to grant the better epistemic position of a particular group (e.g., women) means that the requirements for group membership (here, being a woman) are identified by the very social and/or natural facts on which this group's better epistemic reliability, privilege, or advantage is based. Even now, these social and/or natural facts inevitably single out the essence of "woman".

A more interesting attempt to answer the challenge of essentialism has been proposed by Sandra Harding (1993). Her idea is that a marginalised and/or subordinated group should not be regarded as more epistemically reliable or privileged but should simply be considered as more epistemically *advantaged*, in the sense that it gives rise to better hypotheses, accounts, or explanations in the natural and social sciences (Jaggar 1983):

> the activities of those at the bottom of such social hierarchies can provide starting points for thought – for *everyone's* research and scholarship – from which humans' relations with each other and the natural world can become visible. (Harding 1993, p. 54)

In other words, the experiences of marginalised and/or subordinated groups

> provide the scientific problems and the research agendas – not the solutions – for standpoint theories. Starting off thought from these lives provides fresh and more critical questions about how the social works than does starting off thought from the unexamined lives of members of dominant groups. (Harding 1993, p. 62)

With this move, Harding shifts the attention from the context of justification to the context of discovery and stresses the importance of marginalised and/or subordinated standpoints in discerning new problems and in setting novel research agendas. Thus, marginalised and/or subordinated groups (e.g., women) have something important to say regarding how scientific questions are posed and how research strategies are established and pursued, but this does not mean that their standpoints are more reliable or privileged for determining the justification of any given knowledge claim. We feel that such a strategy is untenable.

A shift from the context of justification to the context of discovery does not appear to offer a novel way to efficaciously rebut the objection from essentialism.

It is crucial to consider that we cannot ground the epistemic advantage of a marginalised and/or subordinated group without understanding its location within society and explaining why some social and/or natural facts about this social location, once rightly appreciated, confer an advantage to those who occupy it with regard to the posing of new questions and developing original scientific theories and hypotheses. Again, this means that these social and/or natural facts inevitably single out the essence of "woman"; in general, one cannot confer epistemic advantage to the standpoint of a given marginalised and/or subordinated group without essentialising the social and/or natural facts about its social location. Otherwise, one can refuse to base the epistemic advantage on any social and/or natural fact. Consequently, all marginalised and/or subordinated social groups must be considered epistemically advantaged to an equal extent, and their knowledge of the world must stand at the very same level. Both perspectives, however, are problematic for epistemologies and epistemologies of the sciences.

Moreover, even if Harding's way of articulating her standpoint theory represents a valuable effort to scrutinise the context of discovery and establish an efficacious feminist methodology of discovery for the natural and social sciences (Harding 2009), we should stress that her view does not represent a genuine (i.e., a normative) epistemology. Despite some controversy, it is still commonly held, at least among some scientists and philosophers of science, that a new scientific theory or hypothesis becomes knowledge only when it has been justified, i.e., when it is tested by a scientific community and corroborated within the so-called context of justification. In other words, the normative context of justification must be considered separately from that of discovery and also from other descriptive contexts, such as the context of decision (Amoretti and Vassallo 2010b).[5] If one acknowledges this general tenet and simultaneously rejects the idea that the standpoint of women is more epistemically reliable or privileged in the context of justification – for merely recognising that it is advantaged only in respect to the posing of new questions and formulating fertile theories and hypothesis in the context of discovery – then one is compelled to deny that the standpoint of women could play any special role and

[5] We recognise that the divide between the normative context of justification and the descriptive context of discovery (possibly refined to distinguish other contexts, such as that of decision) is problematic and should be scrutinised in more detail. The issues involved here warrant another paper. For this reason, the argument assuming this divide has been formulated in merely conditional terms. Nevertheless, some points are worth emphasising (Amoretti and Vassallo 2010b). First, affirming that feminist standpoint epistemologists see their claims regarding discovery as relevant to justification does not undermine the distinction between a descriptive context of discovery and a normative context of justification. Second, appealing to purely descriptive notions to define normative concepts, such as that of justification, does not rule out the possibility of characterising the context of justification as separate from that of discovery. Third, denying the divide between a descriptive context of discovery and a normative context of justification implies, at least implicitly, the rejection of the distinction between psychology and epistemology and the embrace of a radical form of naturalism that, giving up the notion of justification, is no longer a genuine epistemology.

have any particular significance in the production of new *justified* scientific theories or hypotheses. In this way, standpoint theories cannot be configured as genuine (i.e., normative) epistemologies because they have nothing to state about the justification of any type of knowledge claim.

We now consider another, perhaps more compelling, attempt to rebut the argument from essentialism. Alison Wylie, for instance, explicitly insists that feminist standpoint theories "must not presuppose an *essentialist* definition of the social categories or collectivities in terms of which epistemically relevant standpoints are characterized" (Wylie 2003, p. 28). According to her, standpoints must be understood as historical and/or sociological constructions that contingently and only with respect to particular and circumscribed epistemic projects may confer better epistemic reliability, privilege, or advantage to those who actually achieve them. In her own words:

> some *standpoints* (as opposed to *locations*) have the especially salient advantage that they put the critically conscious knowledge in a position to grasp the effects of power relations on their own understanding and that of others. The justification that an appeal to standpoint (or location) confers is, then, just that of a nuanced, well-grounded (naturalized) account of how reliable particular kinds of knowledge are likely to be, given the social conditions of their production. (Wylie 2003, p. 34)

Similar arguments have also been developed by Sandra Harding (1986, 1998), Sharon Crasnow (2008), and Kristina Rolin (2009). To summarise, conceiving standpoints as contingent historical and/or sociological constructions, there is no more reason to accept the existence of any fixed, necessary, and general group membership and gender identity criteria that would be able to grant (whether automatically or not) better epistemic reliability, privilege, or advantage. If these considerations are correct and it is possible to deny the need for any fixed, necessary, and general group membership and gender identity criteria, then it appears that there are no further serious grounds to hold that standpoint theories inevitably yield essentialism. Such an objection appears to be reasonable, but further considerations have to be advanced.

First, let us assume that one does not make use of any fixed, necessary, and general group membership and gender identity criteria; thus, one is not committed to essentialise the social and/or natural facts regarding the location of a particular marginalised and/or subordinated group (e.g., women). Nevertheless, to grant the better epistemic reliability, privilege, or advantage of the group "women", one must not only refer to some social and/or natural facts concerning women themselves, but also – and more importantly – valorise, abstract, and generalise the social and/or natural facts in question in such a way that even if one does not essentialise them, one should always idealise them. Unfortunately for standpoint theorists, as Bat-Ami Bar On correctly notes:

> the kind of idealisation that is entailed by valorisation is problematic because rather than working from a conception of practices [we can read it as 'social facts'] as heterogeneous, it includes some while excluding others, presupposing that there are practices that in one way or another are more authentically expressive of something about the oppressed group. (Bar On 1993, p. 92)

To put it another way, if one wants to derive better epistemic reliability, privilege, or advantage from the contingent social and/or natural position of a group, it is necessary to attribute more value to some, but not others, of women's personal experiences, to transform these experiences into a source of epistemic insight, and to abstract and generalise them rather than recognising them as the individual and personal experiences of particular women. For instance, construing women's standpoint on the nurturing tasks that women perform on a daily basis (taking care of the family, cooking, cleaning, etc.) leads to the valorisation of certain dispositions, such as caring or empathy, and then (willingly or not) to the generalisation and idealisation of them as the "right" dispositions of women, those any "normal" woman should have. Again, construing the standpoint of African American women as their social role of caring for the members of extended families, friends, and neighbours leads to the valorisation of certain abilities, such as skill in community building, and then (willingly or not) to the generalisation and idealisation of these abilities as the "right" abilities for African American women, those any "normal" African American woman should have. Thus, social and/or natural facts must be considered normative requirements that convert any difference into deviance and create unwelcome conditions of exclusion or, at least, inferiority. We may be able to discard essentialism, but not its negative and dangerous consequences. Conversely, if one chooses to deny the opportunity to valorise, abstract, generalise, and thus idealise the social and/or natural facts regarding the location of a particular marginalised and/or subordinated group, then she must also give up the possibility of providing a basis for the group's better epistemic reliability, privilege, or advantage. If such a group is not more epistemically reliable, privileged, or advantaged than other groups, then there is no plausible reason to consider its knowledge more epistemically reliable, privileged, or advantaged than that of other groups.

Second, if standpoints are contingent historical and/or social constructions, then different standpoints (such as, on the one hand, the standpoint of a marginalised and/or subordinated group and, on the other hand, the standpoint of a dominant group) cannot be compared with one another because there is no common ground for such a comparison, no fact of the matter independent of the historical and/or social beliefs that shape standpoints themselves. However, if these standpoints cannot be compared with one another, then it is quite absurd to argue that one of them (e.g., that of the marginalised and/or subordinated group) is more epistemically reliable, privileged, or advanced than others; if different standpoints cannot be compared, then they cannot can be ranked. Consequently, the only viable alternative is to renounce the notions of better epistemic reliability, privilege, or advantage and to coherently maintain that all standpoints are equally reliable, privileged, or advantaged. Such a conclusion, unfortunately, is not welcomed among epistemologists and epistemologists of the sciences. Moreover, a similar naturalised epistemology provides an exclusively descriptive account of knowledge and scientific knowledge; it configures itself as a merely scientific and not genuine (i.e., normative) epistemological enterprise. This means that standpoint epistemologies can no longer be regarded as genuine epistemologies (Amoretti and Vassallo 2010b).

3 Concluding Remarks

Standpoint theories have several merits but make a serious mistake in defending the idea that some standpoints are more epistemically reliable, privileged, or advantaged than others. We have shown that defending the better epistemic reliability, privilege, or advantage of a particular standpoint implies that these notions are based on social and/or natural facts, thus yielding essentialism or at least (in the best scenario) its negative and dangerous consequences, such as conditions of exclusion or inferiority. First, those who insist that merely occupying a specific social position is neither necessary nor sufficient to achieve a particular standpoint make no serious inroads against essentialism. Second, shifting from the context of justification to that of discovery or instead conceiving standpoints as contingent historical and/or social construction results in losing the notions of better epistemic reliability, privilege, or advantage (thus regarding all standpoints as equal) or, alternatively, losing the very possibility of characterising standpoint theories as genuine (i.e., normative) epistemologies. Both options are undesirable.

We believe that the obvious solution is simply to abandon the very notion of standpoint together with the idea of formulating a good feminist standpoint epistemology. In fact, we do not need standpoints to recognise the social situatedness of knowledge, the presence of perspectival biases, and the relevance of epistemic dependence, or to emphasise the importance of pluralism and having more democratic and less sexist practices in the sciences. We do not need to suppose that some perspectives are more reliable, privileged, or advantaged than others because it is the very presence of various, and perhaps even conflicting, perspectives on the world that democratises the epistemic enterprise and may eventually yield more objective knowledge (Longino 1990, 2001). Of course, one may object that such a move would probably imply renouncing the characterisation of epistemology in a specific feminist sense (Crasnow 2008). This may be true, but assuming uncritically that a specific feminist epistemology is actually possible may appear to be another way to defend the possibility of individuating and selecting an unambiguous set of social and/or natural traits that would inevitably single out the essence of "woman".

Acknowledgement Maria Cristina Amoretti gratefully acknowledges financial support for her research from the European Grant POSDRU/89/1.5/S/63663. The authors wish to thank an anonymous reviewer for his/her helpful comments and objections.

References

Amoretti, M. C., & Vassallo, N. (2010a). Do feminist standpoint epistemologies of the sciences answer the charge of essentialism? In M. De Caro & R. Egidi (Eds.), *Architectures of theoretical and practical knowledge: Epistemology, agency, and sciences* (pp. 89–102). Roma: Carocci.

Amoretti, M. C., & Vassallo, N. (2010b). *Piccolo trattato di epistemologia*. Torino: Codice Edizioni.

Amoretti, M. C., & Vassallo, N. (2011). On essentialism: Standpoint epistemologies and their unavoidable problem. In Q. Zhou (Ed.), *Applied social science, ICASS-2011* (Vol. 1, pp. 377–382). Newark: Information Engineering Research Institute.

Amoretti, M. C., & Vassallo, N. (2012). On the independence of the social and situated dimension of scientific knowledge from the notion of standpoint. In S. Knauss, T. Wobbe, & G. Covi (Eds.), *Gendered ways of knowing in science: Scope and limitations* (pp. 57–74). Trento: FBK Press.

Bar On, B.-A. (1993). Marginality and epistemic privilege. In L. M. Alcoff & E. Potter (Eds.), *Feminist epistemologies* (pp. 83–100). New York: Routledge.

Baron-Cohen, S. (2003). *The essential difference: Men, women and the extreme male brain*. New York: Basic Books.

Bowell, T. (2011). Feminist standpoint theory. In *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/fem-stan/. Accessed 3 July 2012.

Brooks, A. (2007). Feminist standpoint epistemology: Building knowledge and empowerment through women's lived experience. In S. N. Hesse-Biber & P. L. Leavy (Eds.), *Feminist research practice: A primer* (pp. 53–82). London: Sage.

Collins, P. H. (2000). *Black feminist thought. Knowledge, consciousness, and the politics of empowerment*. London: Routledge.

Crasnow, S. (2008). Feminist philosophy of science: "Standpoint" and knowledge. *Science and Education, 17*, 1089–1110.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilige of partial perspective. *Feminist Studies, 14*(3), 575–599.

Harding, S. (1986). *The science question in feminism*. Ithaca: Cornell University Press.

Harding, S. (1991). *Whose science? Whose knowledge? Thinking from women's lives*. Ithaca: Cornell University Press.

Harding, S. (1993). Rethinking standpoint epistemology: "What is strong objectivity"? In L. M. Alcoff & E. Potter (Eds.), *Feminist epistemologies* (pp. 49–82). New York: Routledge.

Harding, S. (1995). "Strong objectivity": A response to the new objectivity question. *Synthese, 104*(3), 331–349.

Harding, S. (1997). Comment on Hekman's "Truth and method: Feminist standpoint theory revisited": Whose standpoint needs the regimes of truth and reality? *Signs, 22*(2), 382–391.

Harding, S. (1998). *Is science multicultural? Postcolonialisms, feminisms, and epistemologies: Race, gender, and science*. Bloomington: Indiana University Press.

Harding, S. (Ed.). (2004). *The feminist standpoint theory reader*. London: Routledge.

Harding, S. (2009). Standpoint theories: Productively controversial. *Hypatia, 24*(4), 192–200.

Hartsock, N. (1983). The feminist standpoint: Developing the ground for a specifically feminist historical materialism. In S. Harding & M. B. Hintikka (Eds.), *Discovering reality* (pp. 283–310). Dordrecht: Kluwer.

Hekman, S. J. (1997). Truth and method: Feminist standpoint theory revisited. *Signs, 22*(2), 341–365.

Jaggar, A. M. (1983). *Feminist politics and human nature*. Totowa: Rowman and Allanheld.

Kourany, J. A. (2009). The place of standpoint theory in feminist science studies. *Hypatia, 24*(4), 209–218.

Landau, I. (2008). Problems with feminist standpoint theory in science and education. *Science and Education, 17*, 1081–1088.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton: Princeton University Press.

Longino, H. E. (2001). *The fate of knowledge*. Princeton: Princeton University Press.

Lugones, M., & Spelman, E. (1986). Have we got a theory for you! In M. Pearsall (Ed.), *Women and values* (pp. 19–31). Belmont: Wadsworth.

Mackie, P. (2006). *How things might have been: Individuals, kinds, and essential properties*. Oxford: Oxford University Press.

Potter, E. (2007). *Feminism and philosophy of science. An introduction*. London: Routledge.

Robertson, T. (2008). Essential vs. accidental properties. In *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/essential-accidental/. Accessed 3 July 2012.

Rolin, K. (2009). Standpoint theory as a methodology for the study of power relations. *Hypatia, 24*(4), 218–226.

Rose, H. (1983). Hand, brain and hearth: A feminist epistemology for the natural sciences. *Signs, 9*, 73–90.

Stoljar, N. (2000). Essensialism. In L. Code (Ed.), *Encyclopedia of feminist theories* (pp. 177–178). London: Routledge.

Stone, A. (2004). Essentialism and anti-essentialism in feminist philosophy. *Journal of Moral Philosophy, 1*(2), 135–153.

Witt, C. (2011). *The metaphysics of gender*. Oxford: Oxford University Press.

Wylie, A. (2003). Why standpoint matters. In R. Figueroa & S. Harding (Eds.), *Science and other cultures: Issues in philosophies of science and technology* (pp. 26–48). New York: Routledge.

The Democratic Control of the Scientific Control of Politics

Matthew J. Brown

Abstract I discuss two popular but apparently contradictory theses:

T1. **The democratic control of science** – the aims and activities of science should be subject to public scrutiny via democratic processes of representation and participation.
T2. **The scientific control of policy**, i.e. **technocracy** – political processes should be problem-solving pursuits determined by the methods and results of science and technology.

Many arguments can be given for (T1), both epistemic and moral/political; I will focus on an argument based on the role of non-epistemic values in policy-relevant science. I will argue that we must accept (T2) as a result of an appraisal of the nature of contemporary political problems. Technocratic systems, however, are subject to serious moral and political objections; these difficulties are sufficiently mitigated by (T1). I will set out a framework in which (T1) and (T2) can be consistently and compellingly combined.

1 Introduction

The relationship between science and democracy has been of increasing concern to a variety of fields, including STS, policy studies, environmental studies, and philosophy of science. There are a variety of issues and approaches, but there are two main lines of concern: first, whether and in what sense science is or ought to be political – especially whether it ought to be democratized; second, determining

M.J. Brown (✉)
Center for Values in Medicine, Science, and Technology, The University of Texas at Dallas, 800, W. Campbell Rd, JO 31, Richardson, TX 75248, USA
e-mail: mattbrown@utdallas.edu

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_39,

the role of experts in democratic society – how to deal with their authority versus democratic equality, how to render their role more productive and reliable. My goal will be to explore a way these two lines might converge.

I will consider two theses that are each frequently defended and individually compelling (though by no means uncontroversial) but apparently at odds:

T1. **The democratic control of science** – some of the aims and activities of science should be subject to public scrutiny via democratic processes of representation and participation.

T2. **The scientific control of policy**, i.e. **technocracy** – political processes should be problem-solving pursuits determined by the methods and results of science and technology.

I will not attempt to satisfactorily argue these theses independently, though I review some prominent defenses of them, hopefully demonstrating their plausibility. It is enough for my purposes that there is significant interest and support in these claims to wonder about whether they are consistent. Despite the tension between the two – (**T1**) points to an increasing role of the non-expert public, while (**T2**) points to an increase in expert control – they can be combined in a coherent way. I propose that we can make sense of this combination by treating science and politics as parallel and mutually involving processes. I will sketch a framework for such an understanding science and politics, which I will call "*democratic technocracy*."

2 Why Democratize Science?

There are many arguments for increasing democratic participation in science, especially those areas of science that have an impact on politics and public life. In none of those arguments does "democratizing science" amount to simply replacing evidence with votes.[1] These arguments include purely epistemic arguments, from those depending on purely formal results like the Condorcet Jury Theorem or Diversity-Trumps-Ability Theorem, to Mertonian or pragmatist arguments that democracy is a fundamental requirement of the epistemic structure of science. Rather I will emphasize two ways that we can show the need for democratizing science: based on the social status and role of science and based on the role of values in science.

According to the first type of argument, our current apolitical image of science accords it a high degree of *both* social *authority* and social *autonomy*.[2] A conflict arises when according any institution *both* authority and autonomy to such a great

[1] Anderson (2007) covers several of these sorts of arguments.

[2] Because the focus is on science in its role in the public, especially policy, and not in the abstract, what is at issue cannot be merely epistemic authority, if that is understood in a way that is irrelevant to social authority.

degree. As Heather Douglas puts it, “[A]n autonomous and authoritative science is intolerable… A fully autonomous and authoritative science is too powerful, with no attendant responsibility” (Douglas 2009, pp. 7–8). An authoritative institution compels respect or exercises power over some aspect of social life, while an autonomous institution is not influenced by or responsible to anything beyond its own internal norms. An institution that is both authoritative and autonomous creates an unacceptable tension for a democratic society, which is apparent from the types of institutions ordinarily have these roles.

Social authority is a feature of public institutions, such as legislatures or the police; in democratic societies, the legitimacy of that authority depends *inter alia* on that institution being democratically representative, authorized, and accountable. These democratic responsibilities may take many different forms, but an authoritative institution cannot have legitimacy without them.[3] On the other hand, social autonomy is a feature of private pursuits, traditions, or ideologies, so long as they do not cause harm to non-members or the public interest. The only autonomous sphere is the private one, and private beliefs, practices, or associations do not have any special authority in a democratic society.

The analysis of authority and autonomy thus depends on the distinction between public and private. Following John Dewey in *The Public and Its Problems* (1927), an issue is *public* if it has significant consequences for people beyond those directly involved in and responsible for it; it is private otherwise. A more contemporary term for such consequences is *negative externalities*. We can say that *matters of public interest* arise as groups of people are impacted by the consequences of activities in which they do not participate, recognize those effects, and articulate them as such. The impacted group we call a *public*. By contrast, purely private concerns only affect those who are direct parties to the activity.

By definition, if a practice or institution is socially/politically *authoritative* in some realm, then it has consequences beyond those who are engaged in the practice or constitute the institution; it is a *matter of public interest*. Such practices or institutions *should not be* autonomous, at least in a democratic society, because they will then be immune from the sorts of checks that give their authority *democratic legitimacy*. It is a minimal requirement in democratic societies that the affected parties have a voice on matters of public interest.

Thus, the attempt to combine authority and autonomy in our treatment of science creates a serious conflict. As Douglas points out, those who have responded to that conflict (e.g., Feyerabend, the sociology of scientific knowledge) have tended to challenge the “most obvious” part of this tension: *authority* of science (2009, p. 8).[4] Challenging the authority of science amounts to weakening or denying the existence of expertise in politics. This requires us to give up tools in policy-making that we

[3] The type of “authority” in question concerns the voice that experts *qua* experts have over and above ordinary citizens in policy deliberations. The authority of those policies, once adopted, is a separate issue.

[4] E.g., Feyerabend, *Against Method*, (1975, p. 299).

cannot do without, and, given the remarkable success of science, seems absurd in any case (Douglas 2009, p. 8). While science studies work concerned with this tension has been unduly trying to undermine the *credibility* of scientific experts, they should have been questioning the "*legitimacy* of existing norms of cooperation" between experts and the public (Bohman 1999, p. 591). Challenging the autonomy of science amounts to requiring that science be responsive to and guided by public interests and recognize its democratic obligations.

A second approach to the democratization of science comes from the *value-ladenness of science* defended by feminist philosophers of science among others. It is increasingly difficult to deny that social values necessarily play a role in scientific activity at some level. Values might enter in to several phases of scientific inquiry: e.g., choice of research agenda, methodology, proposal of hypotheses, testing and confirmation, or application. Various theorists have given accounts of the way values work in each stage. For example, Kitcher (2001) focuses on the way that values ought to guide the research agenda of science, determining which projects are significant and ought to be prioritized. Douglas (2000) focuses instead on the role of values in validation of theories and hypothesis, specifically on the role they play in guiding decisions about uncertainty (such as what false positive and false negative rates to accept). Kourany (2010) gives an argument grounded in feminist philosophy of science for strong ethical standards and social responsibilities in every aspect of science. Longino (1990, 2002) is concerned with the role of values in guiding background assumptions and the need for pluralism and critical debate in the social structure of science.

If it is true that values play a necessary role in practice of science, then to the degree that the science has consequences for the public interest, public interests ought to be represented in those value-judgments. It would be inappropriate for scientists as a group to impose their value judgments on the public, in a democratic society, when their value judgments have repercussions for the interests and welfare of the public (Douglas 2005, p. 156). Douglas argues that not only must scientists be explicit about how values are used in making judgments, but also that they must actively democratize their work in a deep way in order to work responsibly.[5]

3 The Scientific Control of Politics

The argument for increasing the role of expert control in politics and policy-making depends on an assessment of contemporary political problems and the way they have been handled in democratic societies. Governing by non-expert opinion doesn't work for contemporary political problems: the problems are too technical, such

[5] Douglas's own approach is largely based on models of participatory democracy and the "analytic-deliberative" model set out in *Understanding Risk: Informing Decisions in a Democratic Society* (Stern and Fineberg 1996). My alternative approach will be laid out in Sect. 4.

that non-expert control is extremely unlikely to provide adequate solutions. Most citizens have only a dim view of what is going on in many of the central political problems of the day. Many current policy proposals are too complex for the public or any non-experts to meaningfully evaluate. One need only listen to commentaries on most major legislation to see that few actually understand the *content* of the proposals in question.

Can representative democracy ameliorate this problem? This is, after all, why we elect representatives, who we are supposed to trust to make these decisions in our stead. They can devote themselves to understanding the issues, with the help of their sizable support staff, and so respond appropriately. In practice, things do not work out so well. From issues of climate and environmental science to medicine and healthcare to economic and monetary policy, many prominent and powerful politicians show themselves to be incompetent to deal with the issues.[6]

Some have gone so far to argue that the reaction of the public and the behavior of politicians on these issues constitute a *failure of democracy*. A1 Gore has said with respect to the policy response to climate change:

> Global warming has been described as the greatest market failure in history. It is also—so far—the biggest failure of democratic governance in history. (Gore 2009, p. 303)

Gore attributes lack of progress he sees towards dealing with the problem of climate change to problems with democracy itself. Environmental scientist James Lovelock has gone even further and suggested that we may need to temporarily suspend democracy to adequately address the problem.[7] If democracy is going to be able to handle the complex and technical problems of contemporary society, its relationship with expertise is going to have to be reconfigured. It no longer seems to be the case that we can rely on non-experts to make the final evaluation in such cases.

The problem is, however, deeper and more fundamental. This is because even political problems that *seem* to be non-technical *actually* require technical expertise for adequate solutions. Indeed, the sort of problems we're more ready to turn over to politics *without* consulting expert opinion may in fact be the most complex and technical. Many of the most controversial political debates are conducted *not* on the basis of clashes of fundamental values, but rather they turn on questions of what will *work*, i.e., the most effective resolution of a problem.

Consider the recent debates about health care policy reform in the United States, which have a long history but have been especially at the forefront of political debate since the debate and passage of the Patient Protection and Affordable Care Act in 2009 and 2010. While there are certainly a number of controversial questions of

[6] Recent exchanges over monetary policy between U.S. Congressman Ron Paul and Federal Reserve chairman and economic expert Ben Bernanke are a particularly evocative version of this. See, e.g., http://www.theatlantic.com/business/archive/2011/07/bernanke-to-ron-paul-gold-isnt-money/241903/

[7] See also Mark B. Brown, "Is Climate Change Good for Democracy," Center for Values in Medicine, Science, and Technology, September 2011. http://www.utdallas.edu/c4v/mark-b-brown-is-climate-change-good-for-democracy/

values with respect to health care policy – is it a right or a private service? how does economic efficiency trade off against the welfare of the disadvantaged? – these are not the top points of controversy in the discussion amongst the public, in the media, and in the political arena. Concerns instead focused on questions like: How much will the reforms increase access? How much will the reforms cost? These are factual questions about the cause and effect relationship between implementing a policy and various results. We want to know, *given* aspects of the problem to be solved, whether the policy will solve it (or make it worse) and to what extent. Many of the questions involve knowledge of the current healthcare system, economics, actuarial science, tax policy, etc.

This is a very general feature of political debate. Without minimizing the importance of conflicts over values, much political controversy turns around complex *factual* questions. On welfare, we wonder whether a policy will spur or discourage job-seeking; whether it will provide enough for the recipients to live on; whether they will be able to game the system. On taxes, whether it will generate enough revenue to cover current spending; whether various groups will pay more or less; whether it will be more efficient. On economic stimulus, whether it will work to bring various economic indicators up in a certain amount of time.

Much political controversy centers around factual questions about whether policies will work to meet stated goals, to solve problems of public interest. But whether some policy will work is not settled by value judgment. Nor does there seem to be compelling evidence that whether a policy will work is well-tracked by public opinion or policymaker judgment. In order to make these determinations, evidence must be gathered and evaluated. Models may need to be constructed, tested, and applied. Consequences may need to be monitored and further revisions considered.

In other words, often what is necessary in political problem-solving is the kind of expertise and inquiry that has proven effective in the sciences: evidence-based, systematic, experimental. This does not necessarily mean that what we need are experienced scientists or technologists – what we need is the same *kind* of expertise but applied to a different subject matter. Policy should be directed by those who are experts *at solving political problems*.[8]

4 Putting the Two Together

The two claims that I have discussed are apparently incompatible:

(T1) tells us that science should be controlled democratically – guided by the public.

[8] Philosophers who have objected to the idea that policy should be directed by experts will be addressed in Sect. 5.

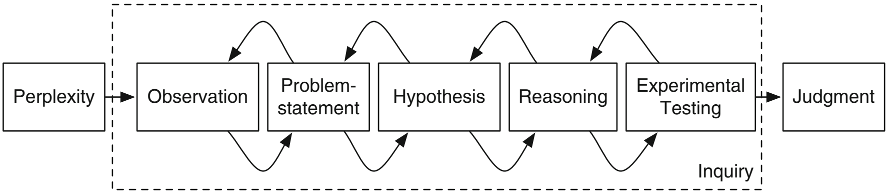


Fig. 1 The pattern of inquiry according to Dewey (Simplified)

(T2) tells us that policy should be decided by expertise and scientific inquiry, not by non-experts – so apparently, not democratically.

The tension arises when our interpretation of (**T1**) is guided by our ordinary conception of democratic politics and (**T2**) by a traditional conception of science. Our ordinary conception of democratic politics puts prime emphasis on public opinion, discussion, and votes. Our ordinary conception of science is technical, value-free, and distant from political engagement.

The two claims can be coherently combined by thinking a little differently about the nature of both science and democracy. As Bohman (1999) says, "both democracy and science must be transformed" in light of their interaction (p. 591). We should regard the *central* process of politics as *inquiry*, in precisely the same sense of 'inquiry' as the central process in science and technology, governed by the same sorts of methods and norms. On the other hand, as explored in Sect. 2, the norms governing science include *not only* considerations of evidence and reasoning, but also democratic and ethical obligations. We scientize political inquiry only after we democratize our conception of scientific inquiry. Call this approach to reconciling these claims *democratic technocracy*.

We can bring out the parallels between science, technology, ethics, and political action by thinking about *inquiry* much as John Dewey did,[9] as an experimental problem-solving process, beginning with a state of perplexity and concluding with a judgment that resolves that perplexity. Inquiry on Dewey's account consists of functionally defined, reciprocally connected phases (Fig. 1). The phases Dewey describes are not surprising or controversial, but what is important is that each phase stands on a par with the other phases as necessary functional components of an recursive process aimed at the resolution of a problem. The upshot is that the adequacy of any component of inquiry faces two tests: compatibility with the other phases of inquiry, and the ability of the whole to produce a judgment that resolves the initial perplexity. Dewey combines these two features under the term "functional fitness" (Dewey 1938, p. 114).

[9] The theory of inquiry was a major concern throughout Dewey's career, including works such as *Studies in Logical Theory* (1903), *How We Think* (1910/1933), *Essays in Experimental Logic* (1916), and *Logic: The Theory of Inquiry* (1938).

In the context of this project, another useful feature of Dewey's account is that the same pattern applies to any type of inquiry: to research in physics, to medical diagnosis, or to choosing a climate mitigation policy. This works because it is a relatively open interpretive framework – it does not overgeneralize from specific features of physics. The framework still has significant normative bite, however. For example, given that evidence is produced as a functional component of an inquiry in the context of solving a particular problem, evidence produced in one inquiry cannot be taken for granted in an inquiry in a very different context. Most "evidence-based policy" guidelines make that mistake, of taking the validity of evidence across contexts for granted; this is one of the (many) reasons such guidelines are incomplete or flawed.

Inquiry of any kind becomes democratized in at least two ways. First, there can be public input into the different phases of inquiry. For instance, there may be situated knowledge that inquirers must aggregate in order to better understand the situation – as when farmers, environmentalists, and those living downstream may have information about the use of a fertilizer in a particular locale that laboratory and field scientists may lack access to. And as in Sect. 2, there is also a major role for public input about value-judgments in the various stages of inquiry. One practical example of such a democratized framework for inquiry is the analytic-deliberative method of *Understanding Risk* (Stern and Fineberg 1996, pp. 16ff, especially p. 28).

This framework shows that democratized inquiry requires a thorough interweaving of scientific-technical experts, political actors, and interested publics. Such analytic-deliberative frameworks may not be feasible or appropriate in all cases. The inquiry may be more removed from the experience of non-experts, may be so technical that the public is unable to engage fruitfully, or they may be long-term inquiries in which direct and meaningful participation is unworkable. In some situations, public participation and deliberation may be counter-productive (Jasanoff 2003). We can democratize inquiry in a second way by having the expert inquirers themselves acting as representatives of the public.

I am not merely suggesting that scientists should act "in the public interest." Indeed, in this simple formulation, many would say this is precisely what scientists do: by engaging in pure research, scientists act in the public interest in advancing and communicating knowledge. Rather, it means that inquirers in democratized inquiry have the same responsibilities to the public as other public officers: legislators, judges, police, bureaucrats, etc., i.e., the responsibility to democratically represent the public, a complex set of activities including, "authorization, accountability, participation, deliberation, and resemblance"(Brown 2009, p. 8).

For example, scientists could be *authorized* by bodies like the National Academy of Sciences but in ways that would assess not only their technical proficiency but also their social responsibility. Their work could be held more *accountable* if they had to make explicit the role of values in their decision-making for scientific debate and public scrutiny (Douglas 2009, p. 173). And they could help increase *resemblance* by ensuring that the scientific community does not systematically exclude any demographics in society or simply by consciously and explicitly considering a variety of social perspectives (Brown 2009, pp. 228–231).

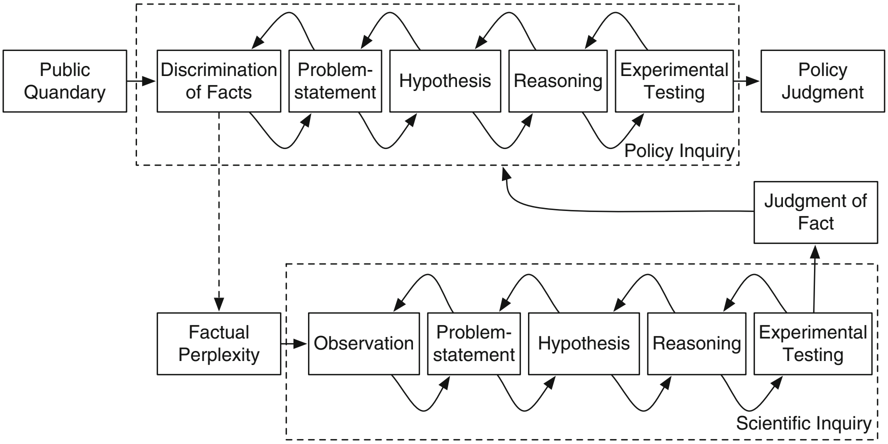


Fig. 2 Policy inquiry spurring scientific inquiry de novo

Turn now to the case of policy inquiry, which I have suggested can be captured by the same pattern of inquiry. The *perplexity* that spurs the inquiry in that case is a *public* quandary, as opposed to a merely private issue. The perplexity is a *matter of public interest* that must be articulated in a participative and democratic process. Policy-making is a response to such problems, a form of inquiry aimed at their resolution. It may be one in which integrating competing value-claims is as important as determining the facts, but all the same it is a form of inquiry.

Policy inquiry remains a rather broad category. In some cases, relatively unstructured and ad hoc public groups can engage in cooperative inquiry leading to a policy judgment. But in the sort of political problems and public quandaries discussed in Sect. 3, much more structured and systematic approaches are necessary, including reliance on the organized institutions of science and government. A central role must be played by a new form of expert: experts at conducting policy inquiry. At the same time, a large role remains for the public – but the same sort of role imagined for the public in the case of science that bears on matters of public interest. In many cases, scientific experts will also play a role in cooperation with policy experts and the public.

In the case of democratized policy inquiry, perplexities of fact may arise that require scientific inquiry *de novo* (Fig. 2). Indeed, the need for gathering new evidence, solving new problems about what is going on and *what* causal structures exist that can be made use of is a pervasive need in modern political practice.

While sometimes knowledge exists prior to the policy inquiry in a pre-packaged form, in general, the political context frames new scientific inquiries. Because inquiry is a contextual problem-solving process, this framing is the only guarantee that the results of scientific inquiry will be relevant and adequate to the political task.

In the case of scientific inquiry spurred and framed by a policy inquiry, we can see how this model treats science and policy-making as both mutually involving and also parallel processes. Both follow the same basic pattern of inquiry. Policy inquiry not only makes use of the results of past inquiries and the methodological lessons of scientific inquiry, but also may spin off scientific inquiries that can respond directly to the problems of fact that it raises. These inquiries are not only framed by the policy issue, but they must be democratically responsible in precisely the same ways that policy inquiries must be.

Jasanoff (2009) spoke hopefully of the place of science in the new administration in the U.S. that it might accept "the essential parallelism between scientific learning and democratic learning." The framework of *democratic technocracy* provides a way of recognizing that parallelism and resolving many of the difficult problems where science and democracy meet.

5 The Threat of Technocracy Ameliorated

There are many objections to technocratic governance that have made it seem an unpalatable response to the sort of problems raised in Sect. 3, and which may be taken to cast doubt on *democratic* technocracy. To the contrary, the framework sketched here ameliorates all of the serious problems with technocracy.

First, technocracy is associated with "The pursuit of technical perfection for its own sake" (Mitcham 1997, p. 263). A common theme among philosophical critics of technology, coming from otherwise diverse points of view, is that the increase of technology brings along with it a focus on efficiency or instrumental rationality to the exclusion of all human values and ends. It should be clear that these sorts of problems do not apply to the framework of democratic technocracy. Democratic technocracy begins with quandaries that are *matters of public interest*, not with purely technical problems or problems defined by the experts.

A second problem with technocracy is a result of the special status accorded to experts. Expert rule as traditionally conceived confronts the problem of the experts themselves ceasing to be agents of the common good and instead becoming a distinct ruling class serving their own interests. Dewey (1927) was concerned to combat this form of bare technocracy (see especially pp. 364–5). These are precisely the problems that the *democratization* of technocratic inquiry is meant to solve. According to the framework of democratic technocracy, every kind of inquiry with public ramifications must be democratized, must involve either public input into the stages of inquiry or democratic representation on the part of the inquirers. Furthermore, the democratic obligations of inquirers increase as their work becomes more a matter of public interest. In the case of policy inquiry, such obligations are paramount. The proposed democratic interactions and representative obligations would prevent policy experts from becoming a specialized class.

Turner (2001) addresses several political problems of the role of experts in a democratic society, including the idea that experts pose a threat to democracy because “expertise is treated as a kind of possession which privileges its possessors with powers that the people cannot successfully control, and cannot acquire or share in” (p. 123). But this is not the case in democratic technocracy’s conception of expertise. According to the account laid out above, policy experts must be accountable in that they must inquire into quandaries that are genuinely matters of public interest, their value judgments must be subject to public input and oversight, and the public must even share in the production of knowledge and policy where doing so will help solve problems more effectively. Both scientific and policy experts must be accountable and responsible in these ways in part to avoid Feyerabend’s worry that “science education” become “a form of state propaganda” (Turner 2001, p. 124; Feyerabend 1978, pp. 73–76).

Finally, a common response to technocratic governance is that it overestimates the power of expertise and scientific inquiry to manage complex social systems. But if we do not use the knowledge and methods of our most powerful tools of inquiry to control these complex systems, what shall we do instead? Leave it up to haphazard fortune? To public opinion? In the face of public problems, we can either do nothing (on the conservative principle that any attempt to fix things is likely to make it worse) or we can try to do something to ameliorate the problem. If we choose the latter, then we should use all of the resources of intelligence at our disposal, including the knowledge and methods of science and technology.

6 Conclusions

I have argued that two significant and often defended (if controversial) theses – (**T1**) that science ought to be controlled and accountable to the public and (**T2**) that policy ought to be controlled by the methods and results of science and technology – can be coherently combined into a compelling framework. This requires transforming our understanding of science to be a value-laden practice and of politics to be a form of problem-solving inquiry, a view I have called *democratic technocracy*. I elaborated the view by connecting it with Dewey’s theory of *inquiry* – the common denominator between science and politics.

This essay leaves open many pressing issues of the relation of science to democracy. It does not begin to address, for example, the question of which types of public participation and deliberation (consensus conferences? citizens’ juries?) serve to help or to hinder scientific expertise in policy-making. Instead, it addresses a fundamental question about the relation of science and democratic politics that lie at the root of such questions. Nevertheless, the framework I have provided for understanding that relation significantly reorients thinking about these issues with many concrete ramifications for specific issues.

We need to think about the jobs of scientists and policy-makers as overlapping, rather than wholly distinct ones to be treated separately by the policy process. Policy-makers ought to be a kind of technical experts, proficient at directing policy inquiry and bringing effective judgment to public quandaries. On the other hand, we should recognize that scientists (potentially) have responsibilities as representatives of the public (Brown 2009, pp. 14, 259). Science should be thought of as a public trust (Dewey 1939, p. 170), not just in the sense that in many places, most scientists are professors at public universities, but in the sense of expressly pursuing the public good and being public accountable for it.

Of course, these points go not only for the work of policy-relevant scientific experts, but also for the parallel work of policy experts. We should think of policy as Dewey did, as an experimental, cooperative inquiry aimed at resolving problems of public interest. While real conflicts will arise, and the need for "politics" in the traditional sense will never go away, shifting the center of gravity of policy-making away from the clash of ideology and public opinion toward the cooperative enterprise of solving shared problems may help to resolve pressing contemporary problems of science and politics.

References

Anderson, E. (2007). The epistemology of democracy. *Episteme: A Journal of Social Epistemology, 3*(1), 8–22.

Bohman, J. (1999). Democracy as inquiry, inquiry as democratic: Pragmatism, social science, and the cognitive division of labor. *American Journal of Political Science, 43*(2), 590–607.

Brown, M. B. (2009). *Science in democracy: Expertise, institutions, and representation*. Cambridge, MA: MIT Press.

Dewey, J. (1903). *Studies in logical theory*. Chicago. University. The decennial publications (2d ser., Vol. XI). Chicago: The University of Chicago press.

Dewey, J. (1916 [2007]). *Essays in experimental logic*. Carbondale: Southern Illinois University Press.

Dewey, J. (1927[1986/2008]). *The public and its problems* (The later works of John Dewey, Vol. 2). Carbondale: Southern Illinois University Press.

Dewey, J. (1933 [1986/2008]). *How we think: A restatement of the relation of reflective thinking to the educative process*. (The later works of John Dewey, Vol. 8). Carbondale: Southern Illinois University Press.

Dewey, J. (1938 [1986/2008]). *Logic: The theory of inquiry*. (The later works of John Dewey, Vol. 12). Southern Illinois University Press.

Dewey, J. (1939 [1986/2008]). *Freedom and culture*. (The later works of John Dewey, Vol. 13). Southern Illinois University Press.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science, 67*(4), 559–579.

Douglas, H. (2005). Inserting the public into science. In S. Maasen & P. Weingart (Eds.), *Democratization of expertise? Exploring novel forms of scientific advice in political decision-making* (Sociology of the sciences yearbook, Vol. 24, pp. 153–169). Dordrecht: Springer.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Feyerabend, P. K. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London and Atlantic Highlands: New Left Books.

Feyerabend, P. K. (1978). *Science in a free society*. London: New Left Books.

Gore, A. (2009). *Our choice: A plan to solve the climate crisis*. Emmaus: Rodale.

Jasanoff, S. (2003). Technologies of humility: Citizen participation in governing science. *Minerva, 41*(3), 223–244.

Jasanoff, S. (2009). Essential parallel between science and democracy. *Seed Magazine*. http://seedmagazine.com/content/article/the_essential_parallel_between_science_and_democracy. Accessed 16 June 2012.

Kitcher, P. (2001). *Science, truth, and democracy*. Oxford: Oxford University Press.

Kourany, J. A. (2010). *Philosophy of science after feminism*. Oxford: Oxford University Press.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton: Princeton University Press.

Longino, H. E. (2002). *The fate of knowledge*. Princeton: Princeton University Press.

Mitcham, C. (1997). Engineering design research and social responsibility. In K. Shrader-Frechette & L. Westra (Eds.), *Technology and values* (pp. 261–278). Lanham: Rowman & Littlefield.

Stern, P. C., & Fineberg, H. V. (Eds.). (1996). *Understanding risk: Informing decisions in a democratic society*. Washington, DC: National Academies Press.

Turner, S. (2001). What is the problem with experts? *Social Studies of Science, 31*(1), 123–149.

Harm, Reciprocity and the Moral Domain

Alejandro Rosas

Abstract Can moral norms be unified under one superordinate content, such as harm? Following the discovery that children at an early age distinguish between moral and conventional norms, this question has been the focus of a recent interdisciplinary debate. Influential critics of the moral-conventional distinction have argued that the moral domain is pluri-dimensional and perhaps not even formally unified. Taking the five foundations theory proposed by Haidt and collaborators as guiding thread, I criticize two influential experiments against the unifying role of harm and point to new evidence from psychopathology and cognitive psychology supporting the hypothesis that harming innocent people is a core concern of norms identified as moral independently of cultural settings.

1 Introduction: Does Morality have a Unifying Content?

In the 1980s Turiel and collaborators reported experimental results showing that children about 4 years of age distinguish between moral and conventional rules. Children singled out proscriptions of behaviors involving harm, injustice or violation of rights, and considered them, in contrast to conventional ones, as valid independently of particular authorities, times and places (Turiel 1983). The special design of those experiments was subsequently known as the moral/conventional task and the theory that distinguishes morality from convention as domain-theory. Domain-theory and the moral/conventional task were first meant as a critique of Kohlberg's stages of moral development. In Kohlberg's view, moral understanding develops relatively late after a stage where moral norms are regarded as social conventions, but Turiel found that 4-year old children already distinguish between

A. Rosas (✉)
Philosophy Department, Universidad Nacional de Colombia, Bogotá, Colombia
e-mail: arosasl@unal.edu.co

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0_40,

moral and conventional rules. Subsequent interdisciplinary debate has focused less on developmental stages and more on the nature of moral judgment. The main concern is whether it is possible to fix the characteristics distinguishing moral norms from other types of social norms and to unify the moral domain.

In this paper I shall argue that norms about harming the innocent are indeed a core concern of all norms that are considered moral, independently of cultural setting. Critical to this harm-hypothesis is to avoid formulating the content that triggers the moral attitudes with concepts that express those attitudes as a matter of logic (Stich et al. 2009, p. 95). For example, to use the concept of rights is circular, because a right, at least an inalienable right, logically implies the evaluative attitudes that single out the moral domain. I shall first re-conceptualize in this introduction the contents of the moral domain following Haidt's foundations theory. In the second section I shall criticize Haidt et al. (1993), who claim to prove that the moralization of disgusting behaviors is unrelated to perceptions of harm. In the third section I criticize an attempt by Kelly et al. (2007) to show that harm not always triggers moral attitudes. In the last section I briefly summarize evidence from psychopathology and cognitive psychology that favors the harm-hypothesis.

Cultural psychologists were among the first to criticize the claim that moral norms involve only harm, justice and rights. Critics carried out experiments similar in design to the moral/conventional task, accepting the formal criteria by which the task identifies moral judgments, namely judgments that subjects are prepared to view as legitimately guiding behavior independently of place and authority. Critics of domain-theory reported that subjects from non-western cultures moralize rules that are not about harm or justice. In perhaps the first critical study of this sort, Schweder, Mahapatra and Miller (1987) carried out experiments showing that Indian subjects belonging to the Hindi culture moralize norms about issues of purity, chastity and respect for status. Examples of such norms are prohibitions to eat certain types of food, to eat fish if you are a widow, to marry across levels in the social hierarchy, to have physical contact with others in the menstruating period.

Turiel and followers took these criticisms seriously (Turiel et al. 1987). One type of answer they gave is illustrated by their discussion of the rule prohibiting widows to eat fish. Turiel and collaborators argued that the difference between Hindis and westerners in this case is only apparent and depends on differences in the factual assumptions. It so happens that in the Hindi culture, people believe that the deceased survive spiritually, and that fish is an aphrodisiac. By eating fish, widows will intensify their sexual appetites and eventually offend their husbands, who continue living and feeling, though not in bodily form. Sexual infidelity is something westerners also disapprove of; but most westerners do not believe that deceased spouses can be thus offended. A difference in their background factual beliefs is responsible for the fact that the two cultures entertain different moral views on whether widows are allowed to eat fish. Arguably, both cultures understand infidelity as an offense and an instance of harming and only background factual beliefs make it possible for one culture to perceive harm where another does not.

Nonetheless, Shweder and collaborators argued that the contents of moral norms include a wider set of contents than merely harm and justice. They grouped harm, justice and rights within the domain of autonomy and postulated two further domains: community and divinity. The cultural psychologist Jonathan Haidt has followed this lead. With his collaborators he formulated a theory of five foundations or domains of morality. These five foundations are grounded on five psychological systems that detect and react not only to harm/care and fairness/reciprocity, but also to the three "binding" domains: in-group/loyalty, authority/respect, and purity/sanctity (Haidt and Joseph 2008). Foundations theory contains one novel claim. It goes beyond the analysis of discourse in cultural psychology and postulates five robust psychological systems with innate content that organize the moral world in advance of experience. Those systems have an evolutionary origin and function in a similar way to taste buds. More precisely, they exist as flexible, generative modules, in a prewired rather than hardwired brain (Haidt and Joseph 2008), detecting moral domains and motivating action. Haidt and collaborators have also argued that political liberals base their moral intuitions on the first two foundations, whereas political conservatives generally rely upon all five.

Turiel described the content of morality as concerning harm, justice and rights (Turiel 1983). Because of the threatening circularity addressed above, it is advisable to drop rights and justice from the description of the content of moral norms. Haidt's five foundations describe the content of norms in a way that allows construing eventual connections to universalizability or authority independence as empirical. I propose to reformulate Turiel's hypothesis, understanding harm in a broad sense that includes what Haidt calls harm/care and fairness/reciprocity. The link between harm/care and fairness/reciprocity is not arbitrary. It emerges if "harm" is given the broad theoretical meaning of making others worse off (Royzman et al. 2009, p. 159, 166). The evolutionary theory of cooperation can also help here. Haidt and Joseph (2008) mention that the system detecting and responding to harm/care evolved through kin selection in mammalian maternal care and then spread to non-relatives. This suggests that some early psychological mechanism, e.g. concern for suffering of offspring, spread from cooperation within to cooperation beyond the family. But natural selection will not favor its spread to non-relatives unless it happens on a reciprocal basis, turning concern conditional. When humans started living in larger groups, reciprocal care may have extended to non-relatives in cases of dire need (threats by nature or third parties, injury or sickness) to provide mutual insurance against misfortune. Non-relatives also required mutual assurance that vicinity would not result in episodes of assault and robbery. Hume speculated that before humans could live peacefully and produce goods collectively they first had to develop conventions regarding property, and disown force or fraud as means of transferring possessions. Agreements against mutual aggression would be sustained naturally by trust, which arguably relies on the same mutual concern that made relatives care for each other. Transposing Hume's talk of convention into the evolutionary key, we could say that the pre-wired domains of harm/care and fairness/ reciprocity are deeply intertwined as adaptations for social cooperation between non-relatives. From a psychological perspective, it is likely that we treat being cheated on positive exchanges in a continuous manner with being harmed or suffering aggression.

The previous argument suggests that some resources of the harm/care system flowed into the fairness/reciprocity system, so that both form a continuum negatively reacting to harm in self and others in social interactions. But Haidt's theory includes yet three domains beyond harm and reciprocity. Facing this multiplicity, some philosophers remain skeptical about the possibility of unifying morality under a common content (Sripada and Stich 2006), and others argue outright that all attempts at unification have failed (Sinnott-Armstrong and Wheatley 2011). These denials notwithstanding, there seems to be something like a default philosophical intuition that calls for a unifying principle, such that even skeptics have tried to formulate unifying principles of a formal nature. Sripada and Stich (2006) postulate intrinsic motivation; Nichols postulates a "strong negative affect" (Nichols 2004, p. 64). A deep philosophical prejudice urges us to search for a common factor in all norms labeled *moral*. Moreover, a unifying content would give completion to the more formal proposals relying on intrinsic motivation or strong affect. Content would be what activates similar emotions or motivations.

But as is well known, experimental philosophy has shown how unreliable philosophical intuitions are as guides to what "we" or a majority of language users mean (Knobe and Nichols 2008). So in the next two sections I will examine two experiments that claim empirical evidence against a core role for harm in moral judgments. The first experiment by Haidt et al. (1993) claims that harm does not play in the minds of people the role hypothesized by moral/conventional distinction. It claims that low SES subjects moralize disgusting/ disrespectful behaviors unconnected to harm. These behaviors belong to the purity/sanctity or the in-group/ loyalty system; and the fact that they are moralized shows that those systems manage moral contents independently of the harm/care system. On the other hand, Kelly et al. (2007) examined the other side of the coin, namely whether harming behaviors are always judged as moral transgressions. They claim to have found negative evidence. I shall argue that these two experiments do not provide the empirical evidence they claim to provide.

Intuitively, the "binding" domains of in-group/loyalty, authority/hierarchy and purity/ sanctity have a different content when compared to harm and reciprocity. In the fourth section I endorse a fairly recent hypothesis, backed by empirical evidence, claiming that a moral dyad with a harming agent and a suffering patient actually functions in our minds as a unifying cognitive template for all five moral domains (Waytz et al. 2010; Gray et al. 2012), suggesting that the three "binding" domains are subordinate to the harm and fairness domains.

2 Disgusting/Disrespectful Behaviors and Harm-Based Morality

Haidt and two Brazilian collaborators conducted an experiment designed to show that some people moralize behaviors unrelated to harm or reciprocity rules, because no one is harmed (Haidt et al. 1993). They claim to have found positive evidence, refuting attempts to unify moral issues under the domain of harm. The items chosen were disgusting or disrespectful behaviors performed in private. I think the evidence is not conclusive. The problem arises from a peculiarity of disgusting behaviors, namely, that they are in fact harmful when witnessed. In this section I argue that in the case of low SES subjects who moralize these behaviors, harm is in fact playing a role. In contrast to high SES subjects, they lack the experience of a private sphere in their daily lives, so they cannot easily picture those behaviors as private.

The study presents three disgusting and two disrespectful stories to low and high SES groups in three cities, two in Brazil (Recife and Porto Alegre) and one on the USA (Philadelphia). The results were interesting regarding the three disgusting stories: a man masturbates with a dead chicken and then cooks it for dinner; a family eats their pet dog after it died in an accident; a brother and sister kiss each other passionately on the mouth. Private behaviors of this type were found way more offensive than disrespectful ones. The first thing to point out is that the authors acknowledge that disrespectful behaviors are harmful if they are performed in public. They produce psychological harm: painful emotions in those that perceive them. The authors approvingly paraphrase Turiel:

> . . . burning a flag in public and wearing a bikini to a funeral are not purely conventional violations . . . Given the social significance of these acts, other people will be psychologically harmed, so these acts should be condemned by anyone with a harm-based morality (Haidt et al. 1993, p. 615).

Turning now to disgusting behaviors, contradictory evidence exists presently as to whether they cause harm to witnesses. Nichols (2004) presented data showing that they don't "hurt", but Royzman et al. (2009) found that they do "negatively affect" others. Haidt et al. (1993) include in their study a "bother-probe": "Imagine that you actually saw someone [performing the act]. Would it bother you, or would you not care?" (Haidt et al. 1993, p. 617). The bother probe gives information as to whether subjects would be offended if they happened to witness the behavior in question, and the data support the claim that disgusting behaviors cause psychological harm. The mean percentage of subjects that found the three disgusting stories offensive was 73 % (Haidt et al. 1993, p. 618).

But in their study, they want to prove that some other reason unrelated to harm must account for the moralization of disgusting behaviors when performed privately:

> . . . it is an empirical question whether disgusting acts such as incest are moralized because of their potential for harm, or whether they are considered intrinsically wrong regardless of their consequences (Haidt et al. 1993, p. 615).

Basically, they acknowledge two possible explanations for moralizing those behaviors. One invokes the fact that disgusting behaviors cause harm if witnessed. Since being witnessed is not an intrinsic property of those behaviors, they express this saying that moralization is triggered by their potential for harm. But this is a trap they should not go into. Allowing for potential harm as a ground for moralization makes it very difficult to formulate the alternative hypothesis. This hypothesis says that if the behaviors are performed privately and no harm is done, then moralization has to be explained on grounds other than harm. But if potential harm is a possible explanation, it is open for an advocate of harm-based morality to say that disgusting behaviors performed privately are moralized precisely because of their *potential* for harm. However, their data do not seem to show that subjects moralize on the basis of potential for harm. Rather, I think the design of their experiment raises another issue: low and high SES subjects do not have comparable experiences of privacy.

Before explaining this idea let me first summarize their results. The study shows that high and low SES groups differ in moralizing attitudes towards the disgusting behaviors. But both groups found the three disgusting stories offensive (mean percentage across groups was 73 %, Haidt et al. 1993, p. 618). Also, no large differences were found in the "perception of harm" between low SES groups (mean is 42 %) and high SES groups (mean is 35 %), measured through "yes" answers to the question: "Is anyone hurt by what [the actor] did? Who? How?" (Haidt et al. 1993, p. 618). Regrettably, this is the mean for all five stories. It would have been better to report the values for each story, for they could differ significantly. Regarding moral attitudes, only low SES subjects moralize the disgusting or disrespectful behaviors (the data for moralization attitudes is reported in Tables 1 and 2, pp. 619, 620). High SES groups are permissive. This difference is linked to SES and not to whether the groups actually come from Brazil or the USA.

The explanation they give is that harm is not a necessary condition for moralization of behaviors. However, as noted before, an advocate of a harm-based morality could invoke potential harm and argue that the study shows only that low SES groups condemn on the basis of potential harm, while high SES groups do not. For example, she could claim that disgusting behaviors done in private shape a character that will eventually harm others. Though a plausible explanation, the belief about character-formation would not explain the difference found between low and high SES, because there is no reason to confine it to low SES subjects. Therefore, I do not think that invoking potential harm fits their data.

A more plausible explanation, one that preserves the role of harm in moralization, invokes a different factor, namely, the different experience that low and high SES groups have of spheres of privacy. One of the striking findings of Haidt et al. 1993 is that SES has a very large effect on attitudes towards disgusting and disrespectful behaviors done in private, while a difference between USA and Brazil, or westernization, has a small effect among adults (Haidt et al. 1993, p. 619; 625). Plausibly, SES determines whether an individual enjoys a private sphere or not and how effectively it guarantees purely private acts. Members of low SES groups live in material conditions that make them physically close in a very literal sense. In the low SES groups of the study, it is usual for whole families to sleep and sometimes live in one room. Arguably, only affluent people enjoy individual private spheres. With no privacy, disgusting behaviors are bound to have witnesses. It makes sense, therefore, to suggest that low SES groups moralize disgusting behaviors on the basis of harm, because they cannot easily picture them being performed in private.

3 Are Some Harm-Based Rules Conventional?

Kelly et al. (2007) obtained experimental data showing that some transgressions involving harm and a victim do not evoke the typical moral response pattern in the moral/conventional task: they are judged to be dependent on time, place and authority. In this section I explain their results in a way that preserves the role of harm in the moral/conventional (M/C) distinction. Two reasons, I shall argue, provide the explanation. The first reason concerns their use of mixed-domain situations (Nucci 2001, pp. 95–97) in 6 items of their questionnaire. Mixed-domain situations do not produce the typical pattern of responses in the M/C task. Mixed-domain situations are such that: (1) Two normative domains range over the same action, namely a moral rule and an authority-based rule; and (2) the two rules are in conflict, i.e., they pull agents or respondents in contradictory directions. In situations of this type, respondents perceive the conflict and resolve it variably, sometimes giving priority to the authority rule.

The second reason affects four items in their questionnaire, labeled Whipping/ temporal, Whipping/ Authority, Spanking/ Authority and Prisoner abuse/Authority. These are not the typical transgressions unambiguously "involving a victim who has been harmed, whose rights have been violated, or who has been subject to an injustice". In contrast to the original Hitting and the Hair pulling schoolyard stories, the victims of these transgressions are themselves transgressors. The harm inflicted on them has the quality of punishment. I shall now develop these two reasons in more detail.

The use of mixed-domain situations in Kelly et al. (2007) is particularly important in stories where the person harmed is not a transgressor. These cases are represented in the two typical schoolyard stories about hitting and hair pulling. In their design, the classical stories are presented not once but twice: the original story and then a reframed version of the same story, preceded by an introduction: "Suppose the teacher had said: 'In this school there is no rule against pulling hair (hitting)'." They label these items the Hair pulling/Authority and the Hitting/Authority cases. The reframing is designed to confirm or disconfirm the moral/conventional distinction: confirmation would follow only if there is little variation in the responses to the original and reframed versions. However, 53 % of participants say it is Ok to hit (14 % say it is Ok to pull hair) if the teacher says it is permitted, whereas only 14 % say it is Ok to hit (4 % say it is Ok to pull hair) when the permission is omitted. This result is their toughest challenge to the M/C distinction. But the result is explained away by calling attention to what Turiel and collaborators have called mixed-domain situations.

An example of mixed domain situations is where a father tells his son to steal flowers from the neighbor's garden. Another example is provided by the famous Milgram experiments on obedience to authority (Milgram 1974). A scientist, presumably investigating the impact of punishment on learning, commands participants to give electric shocks (faked) to an innocent person each time she errs (a confederate that simulates pain), contravening a moral rule against harming the innocent. Typical for these mixed-domain situations is a contradiction between rules: a command by a scientist to give electric shocks contradicts a moral rule against harming innocent people.

In the Hitting/Authority and the Hair pull/Authority items the mixed situation is created when a moral transgression is reframed as a story where an authority explicitly permits the transgression locally. This is also the case in four other items in their questionnaire, labeled the Spanking/ Authority, Prisoner abuse/ Authority, Whipping/ Authority and Military training/ Authority stories. The point about mixed-domain situations is that participants feel the pull of two contradictory rules. In the Milgram experiments, participants have sometimes obeyed in high percentages and have sometimes rejected almost unanimously the scientist's command, depending on the experimental conditions (Milgram 1974). Interestingly, the percentages of approval obtained by Kelly et al. in the pulling hair (14 %) and hitting (53 %) at school when the authority permits the behavior differ widely. In itself, this large difference for two very similar behaviors suggests the effect of uncontrolled factors influencing participants' responses.

In the design by Kelly et al. (2007) participants were not asked whether the school authorities could legitimately change the moral rules about hitting and hair pulling. Rather, it was assumed that they had already issued corresponding permissions and that these were local in scope ("at this school it is allowed . . . "). Participants responding that it is Ok to hit when the school authority permits it could simply mean that it does not violate the authority's rule. Understood in this way, the response does not assert that moral rules are authority dependent. The design used in Kelly et al. (2007) contrasts with one used by Nucci (2001) when investigating the M/C distinction in religious children and youths. The question posed was whether it would be wrong or not for religious authorities to change moral rules. This way of asking does not posit a hypothetical situation where two rules apply, i.e., a mixed-domain situation. Rather, participants have to think explicitly about whether an authority has legitimate jurisdiction over a moral rule.

In the typical experiments that confirm the moral/conventional distinction, the transgressions studied were stealing, hitting, calumny and damaging another's property. These are the types of moral transgression used in the experiment about the M/C distinction in religious children. Targets of these transgressions are people depicted as innocent of any previous harm, i.e., genuine victims of an agent harming to obtain some selfish benefit. In contrast, the battery of stories used in Kelly et al. (2007) contains a group of four transgressions where harm is inflicted on a transgressor as punishment. These four stories are the ones labeled Whipping/Temporal, Whipping/Authority, Spanking/ Authority and Prisoner Abuse/Authority. Harm is here directed at people guilty either of reckless behavior (as in whipping sailors when drunk on duty) or of direct harm, as when a child is spanked for repeatedly hitting other children.

There is an obvious difference between harming the innocent as a means to selfish gains and harming those that have harmed others, intended as punishment. Though few would agree that corporal punishment is appropriate, many people endorse it and would express it if an authority sanctions it as legitimate. This conjecture needs experimental confirmation, but it is a reasonable hypothesis that could explain the results in Kelly et al. (2007). The fact that a high percentage of subjects endorse harm as punishment when authorities allow it does not imply that rules against harming the *innocent* are authority-dependent. Acting independently, the two explanations here developed cover seven of the nine stories used in Kelly et al. (2007). I have not addressed people's attitudes towards cannibalism, nor towards slavery. The cannibalism story did not involve victims and served a different purpose in their paper. People's attitudes towards slavery are more difficult to analyze. They are neither a clear case of the prototypical transgression of harming the innocent nor are they a clear case of punishing the guilty.

4 Harm in Moral Cognition and Motivation

A defense of harm as a unifying content need not discover an *objective* link to harm in the three "binding" domains (in-group/loyalty, authority/hierarchy and purity/sanctity), but rather point out that moralization arises when subjects *perceive* harm to self and others as consequence of violating norms from those domains. A researcher may judge that a given rule from another culture or subculture has no connection to harm. However, it is not the researcher's belief, but the belief of the subject who moralizes that counts, i.e., her own perception of harm. These perceptions are highly variable: they will depend on background beliefs or assumptions that vary cross-culturally, or on SES, or on individual variation. The moral dyad hypothesis, i.e., the idea that moral cognition is mediated by a template that combines a perceived intentional agent with a perceived suffering patient, predicts, for example, that a built-in deficit in the ability to perceive other minds as subjects of experience (including the experience of pain) will induce a greater willingness to harm (Gray et al. 2012). Accordingly, a deficit in perception of pain in others, indicated by higher scores in the Levenson's Self-Report Psychopathy scale, are correlated with lower endorsement of norms related to harm and fairness (Glenn et al. 2009). No effect is observed on the other three foundations. However, the Moral Foundations Sacredness Scale (Graham et al. 2009) revealed that higher psychopathy scores are correlated with a higher willingness to violate all five foundations for money (Glenn et al. 2009). These two results suggest that a deficit in the perception of pain in others is correlated with a reduced moral motivation in all five foundations. Why should this be so if those foundations are unrelated to harm?

One way to explain these results is to hypothesize that the three "binding" foundations are subordinated to the harm and reciprocity system. These two domains would constitute the core set of the innate foundations; the binding domains would hold a subordinate role to the domains of harm and reciprocity. We may picture this subordination as in the model envisioned by Sperber (2005), where a core set of innate modules generates (within a culture) a diversity of other modules nested within them.

Additional evidence for a unifying role of harm comes from a study asking participants to rate the wrongness of moral transgressions across the five moral domains and to identify whether a victim was harmed. Transgressions against fairness, ingroup, authority, and purity norms elicited perceptions of victims or sufferers, even among conservatives. Another study compared response times to “harmful” after seeing “unfair”, “disloyal” and “impure”. These words asymmetrically primed “harm” and were not primed by each other. These and other cognitive experiments reported in Gray et al. (2012) certainly provide suggestive evidence that the role assigned to harm in the pioneering experiments by Turiel and followers still deserves a place among scientific approaches to morality.

References

Glenn, A. L., Iyer, R., Graham, J., Koleva, S., & Haidt, J. (2009). Are all types of morality compromised in psychopathy? *Journal of Personality Disorders, 23*, 384–398.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*, 101–124.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality Social Psychology, 65*, 613–628.

Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers et al. (Eds.), *The innate mind* (pp. 367–444). New York: Oxford University Press.

Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, affect, and the moral/conventional distinction. *Mind & Language, 22*(2), 117–131.

Knobe, J., & Nichols, S. (2008). *Experimental philosophy*. Oxford: Oxford University Press.

Milgram, S. (1974). *Obedience to authority*. New York: Harper & Row.

Nichols S (2004) Sentimental rules: On the natural foundations of moral judgment. Oxford University Press, Oxford

Nucci, L. (2001). *Education in the moral domain*. Cambridge: Cambridge University Press.

Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral-conventional distinction. *Cognition, 112*(1), 159–174.

Shweder, R. A., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (pp. 1–83). Chicago: University of Chicago Press.

Sinnott-Armstrong, W., & Wheatley, T. (2011). Moral judgments are not unified: Why moral psychology needs a taxonomic approach. Manuscript under review.

Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers et al. (Eds.), *The innate mind: Structure and contents* (pp. 53–68). New York: Oxford University Press.

Sripada, C. H., & Stich, S. (2006). A Framework for the psychology of norms. In P. Carruthers et al. (Eds.), *The innate mind: Culture and cognition*. New York: Oxford University Press.

Stich, S., Fessler, D., & Kelly, D. (2009). On the morality of harm: A response to Sousa, Holbrook and Piazza. *Cognition, 113*, 93–97.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.

Turiel, E., Killen, M., & Helwig, C. (1987). Morality: Its structure, functions, and vagaries. In J. Kagan & S. Lamb (Eds.), *The emergence of morality in young children* (pp. 155–244). Chicago: University of Chicago Press.

Waytz, A., Gray, K., Epley, N., & Wegner, D. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*, 383–388.

Index

V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, DOI 10.1007/978-3-319-01306-0,