The Philosophy of Science in a European Perspective

Hanne Andersen · Dennis Dieks
Wenceslao J. Gonzalez · Thomas Uebel
Gregory Wheeler *Editors*

New Challenges to Philosophy of Science

Springer

New Challenges to Philosophy of Science

[The Philosophy of Science in a European Perspective, Vol. 4]

Proceedings of the ESF Research Networking Programme

THE PHILOSOPHY OF SCIENCE IN A EUROPEAN PERSPECTIVE

Volume 4

EUROPEAN SCIENCE FOUNDATION

www.pse-esf.org

Hanne Andersen • Dennis Dieks
Wenceslao J. Gonzalez • Thomas Uebel
Gregory Wheeler
Editors

New Challenges to Philosophy of Science

Springer

Editors
Hanne Andersen
Center for Science Studies
Aarhus University
Denmark

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

Gregory Wheeler
Centre for Artificial Intelligence
(CENTRIA)
Department of Computer Science
New University of Lisbon
Portugal

Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA, USA

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

Thomas Uebel
Philosophy School of Social Science
The University of Manchester
United Kingdom

ISBN 978-94-007-5844-5 ISBN 978-94-007-5845-2 (eBook)
DOI 10.1007/978-94-007-5845-2
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013935949

Printed on acid-free paper

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

Table of Contents

PSE Symposium at EPSA 2011: New Challenges to Philosophy of Science

Wenceslao J. Gonzalez

From the Sciences that Philosophy Has "Neglected" to the New Challenges

This fourth volume of the Programme "The Philosophy of Science in a European Perspective" deals with new challenges in this field. In this regard, it seeks to broaden the scope of the philosophy of science in two directions. On the one hand, this book looks for issues in scientific disciplines that have received little attention so far in the mainstream philosophy of science (e.g., design sciences, communication sciences or forensic science). On the other, it addresses new topics in well-established disciplines, seeking novelty from different angles of the philosophical research. To some extent, this volume tries to embrace the somehow "neglected sciences" as well as the new trends in philosophy of science.

I

Historically, some sciences have received more philosophical attention than others, as it is the case of the natural sciences in comparison with the sciences of the artificial. The more intense level of attention could be for several reasons, including a combination of structural elements and dynamic components. Among these reasons: (i) the existence of a well-established field conceived as a model for other sciences, such as physics for the natural sciences (and even for science, in general); (ii) the development of a domain having many epistemological and methodological influences on other scientific subjects, such as biology (mainly through the Darwinian approach); and (iii) the presence of a scientific terrain having a lot of practical consequences, especially by dint of being an applied science (such as economics or psychology).

These reasons, where structural factors have a special weight, make the noticeable relevance of some scientific disciplines for philosophers of science understandable. Meanwhile, there are other reasons, including clear dynamic traits, which can call for preferences among philosophers of science in favor of some disciplines rather than others. Among such reasons let us recall: (a) the novelty of the scientific field, which makes the philosophical-methodological analysis more difficult; (b) the scarcity of explicit influences of the science beyond the "boundaries" of the field; and (c) the strong interweaving with technology, which makes the distinction between the scientific approach and the technological constituent particularly difficult.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_1,

Initially, when the proposal of the Programme "The Philosophy of Science in a European Perspective" was submitted to the European Science Foundation, the emphasis was on "philosophical aspects of those sciences (some new) that have been hitherto ignored in the philosophical literature."[1] Thus, the general label for the fourth year of the Programme – common for the five teams – was "The Sciences that Philosophy has Neglected." It was a way of pointing out the need for more philosophical attention to certain disciplines. And the title of the present book – "New Challenges to Philosophy of Science" – stresses the relevance of addressing issues in "new sciences" or considering different aspects regarding "traditional disciplines", where there are still many important problems to be discussed.

II

Within this general framework, the five teams of the Programme ESF-PSE organized four workshops in 2011. The teams are focused on Formal Methods (Team A), Philosophy of the Natural and Life Sciences (Team B), Philosophy of the Cultural and Social Sciences (Team C), Philosophy of the Physical Sciences (Team D), and History of the Philosophy of Science (Team E). During the four workshops each team had a special interest in making explicit enlargements of the philosophical discussions according to their lines of research during the four workshops:

1. Joint workshop of Teams A and D on "The Philosophy of Computer Science and Artificial Intelligence", organized by Dennis Dieks and Stephan Hartmann, with Gregory Wheeler as local organizer, in Ponta Delgada, Portugal, September 7-9.

2. Workshop of Team B on "Philosophy of Systems Biology", organized by Marcel Weber and Hanne Andersen at the University of Aarhus, Denmark, August 18-20.

3. Workshop of Team C on "The Sciences of the Artificial vs. the Cultural and Social Sciences", organized by Wenceslao J. Gonzalez in Bucharest, with Adrian Miroiu as local organizer, at the National School of Political Studies and Public Administration, Romania, September 15-16.

4. Workshop of Team E on "The Philosophy of the Sciences that Received Philosophy of Science Neglected: Historical Perspective", organized by Thomas Uebel, with Anastasios Brenner as local organizer, at the University of Montpellier, France, November 18-19.

An additional symposium was organized by Raffaella Campaner and Theo Kuipers in Athens, with Maria Carla Galavotti as Chair during the European Phi-

1 Steering Committee, *The Philosophy of Science in a European perspective* Proposal of a "à la carte Programme" to be submitted to the European Science Foundation, 24 February 2006, p. 3.

losophy of Science Association conference, 5-8 October 2011, under the label of "New Challenges for Philosophy of Science". The contributions have been added to the present book in order to complete the picture of the analyses made in the four workshops organized by the teams.

III

Central contents of the four workshops have been reported on the website of the Programme.[2] In the case of the joint workshop of Teams A and D, the papers gave an overview of the present state in "The Philosophy of Computer Science and Artificial Intelligence", with a special focus on quantum information. The workshop was organized with joint morning sessions and parallel sessions in the afternoon. The use of mathematical and logical methods in philosophy and in science was discussed, paying particular attention to computation and quantum mechanics.

During the joint workshop a wide range of modeling techniques was available on display – methods which drew freely from philosophy, physics, philosophy of science, computer science, applied logic, and mathematics. Among the topics: (i) the philosophical and descriptive features of modeling, connected to new modeling techniques that highlighted a key conceptual insight into the problem; (ii) the role played by the group size in whether cooperation, which has suggested a practical (and testable) prescription for negotiating treaties on climate change; (iii) a novel model of learning that highlights an ability of the agent to project itself into future situations; (iv) the suggestion of the extension to classical methods within logic to accommodate quantum systems; (v) a program regarding philosophy of experimentation that draws on models of agent interaction and causality; (vi) the combination of correlation measures, incremental confirmation measures, and causal models to explain the principles of coherent reasoning; and (vii) the possible limits of formal methods to philosophy and to cognitive psychology.

Regarding the workshop of Team B, the interdisciplinarity of the approach merits attention. This event was combined with a workshop on the philosophy of interdisciplinarity, organized by the Philosophy of Contemporary Science in Practice Group of Aarhus University. Thus, the program on the third day of the interdisciplinarity workshop was, at the same time, the program of the first day of the ESF workshop. The corresponding overlap in scientific content of the two meetings was due to the fact that systems biology is usually described as a highly interdisciplinary science. Hence, the discussion moved from the topic of interdisciplinarity to the interdisciplinarity of systems biology and to other topics in the philosophy of systems biology.

---

2 Cf. "The Philosophy of Science in a European Perspective", *http://www.pse-esf.org/ESFworkshops.htm* (access on 11. 12. 2011).

One of the results of Team B was the conclusion that the direct interaction between philosophers of science working *on* systems biology and scientists working *in* systems biology is very fruitful. Some of the recurrent themes during the workshop were causality, modeling, levels of organization, explanation, interdisciplinarity and integrative frameworks. Thus, it was clear that an orchestrated investigation of the new field of systems biology provides not only a deeper understanding of the philosophical issues related to systems biology itself, but also many interesting opportunities for comparison with other fields and thus cast new light on classical discussions in philosophy of science.

According to the guidelines of Team C for the Programme,[3] the workshop on "The Sciences of the Artificial vs. the Cultural and Social Sciences" has tried to enlarge the field with new reflections on the design sciences as well as with new analyses of the social sciences, including the considerations on complexity in the former and on the applied aspects in the latter. In many ways, the novelty regarding the realm of the topics discussed was clear. The contents of the workshop were structured along three main lines: 1) the *Geisteswissenschaften* and the social sciences; 2) from applied social sciences to the sciences of the artificial; and 3) philosophy of the sciences of the artificial. In addition, there was the Junior Meeting that contributed to complete the picture of the topics discussed in the previous sections.

Following the differences and similarities between the sciences of the artificial and the cultural and social sciences, in this fourth year of research new light was shed on relevant aspects: a) historicity and complexity in the design sciences, including their epistemic and methodological specificity; b) the distinction between "human activity" and "human behavior" as a key category for the analysis of complexity and historicity of sciences such as economics; c) the different kinds of complexity available in the sciences of the artificial (structural and dynamic, ontological and epistemological, etc.); d) the need for a coexistence of explanations – oriented towards causes – and reasons or intentions in the social sciences; e) the important role of interpretation and understanding in archeological explanation, which might be extended to the social explanation; etc.

Concerning the workshop of Team E, the historical orientation of the group has faced a particular challenge and opportunity in addressing the overall topic of year 4 of the PSE Programme, which is the consideration of philosophical issues posed by the sciences that philosophy has tended to neglect. A central aim has been to link the existence of many forgotten precursors to the present approaches to the overall theme of the Team, regarding the plurality of philosophy of science traditions in Europe. The focus has been on sciences that were relatively little

3 They are presented in Wenceslao J. Gonzalez, "Trends and Problems in Philosophy of Social and Cultural Sciences: A European Perspective", in: Friedrich Stadler, Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartman, Thomas Uebel, and Marcel Weber (Eds.), *The Present Situation in the Philosophy of Science*. Dordrecht: Springer 2010, p. 236.

discussed in the past, and the workshop has broadened the perspective to take in also a sustained consideration of the French and the Kantian traditions in the philosophy of science.

Hence, the analytic tradition was labeled under the moniker "received" in the title of the workshop. In this respect, the Team E has considered the early attempts at a philosophy of an otherwise neglected science by members of the analytic tradition, and has also paid an extended attention to the specific contribution that French philosophy of science made to these topics. In addition, there was a round-table discussion featuring members of Team E on "The Contribution of Kantian and Neo-Kantian Philosophy to the Philosophy of the Empirical Sciences Other Than Physics". In this way Team E has done justice both to the "first-order" topic of year 4 and the multifaceted nature of the PSE Programme. The outcome of this meeting suggests a more complex view of European philosophy of science than one that is often assumed. Such a more complex view takes into account, on the one hand, those parts that have been integrated into the analytic tradition and, on the other, those sources and debates that have been forgotten or discarded.

Meanwhile the symposium on the "New Challenges for Philosophy of Science", held within the EPSA Conference, assumed that, in recent years, philosophy of science has been widening its range of interests. Thus, it has devoted specific attention to emerging disciplines that had been previously neglected, such as medicine, archaeology, cognitive science, climate and environmental sciences. This has emphasized the relevance of disciplinary specificities for any reflection on methodological and foundational matters. The emerging disciplines call for conceptual and methodological clarification, and hence stimulate and encourage deeper exchanges between the sciences and philosophy of science.

How contemporary philosophy of science interacts with emerging scientific disciplines can be studied through a clarification of crucial concepts such as those of scientific explanation, prediction, reduction, and through studies on multilevel model building. This requires us to consider that, the theoretical issues and the ways they are addressed, are intertwined with the distinctive practical exigencies and application purposes of these disciplines. Thus, the symposium has argued that this focus on the most recent and innovative trends in philosophy of science sheds some light on its future developments and new directions. The symposium started by considering how philosophy of science has become better and better at understanding the actual practice of various scientific disciplines (e.g., brain science, decision science and social science), and how such insights allow to provide methodological suggestions about how these disciplines ought to be practiced.

IV

All in all, it seems clear that this academic project involves a group of scholars trying to make it real in different ways. They require recognition on behalf of those aware of their efforts. In this regard, besides the institutional acknowledgment to the European Science Foundation for its support, there are several people to be mentioned here for their contribution to the Programme "The Philosophy of Science in a European Perspective" during this fourth year.

Firstly, the clear leadership of the Chair of the program, Maria Carla Galavotti, in coordinating this joint venture of the Steering Committee, the Team Leaders and the participants (members of the Programme and external) should be recognized. Secondly, people directly involved in the organization of the workshops, including the Team Leaders and the local organizers, are to be thanked. Thirdly, the commitment of those persons related to internal matters of the Programme, such as Cristina Paoletti and Beatrice Collina, as well as the representatives of the ESF, such as Arianna Ciula and Barry Dixon is to be acknowledged. Last, but not least, many thanks go to the people working on the edition of the volume, including many of the names already mentioned and some other members of the Vienna Circle Institute, such as Robert Kaller.

Ferrol, 24 January 2012

Teams A and D
The Philosophy of Computer Science and Artificial Intelligence

JESSE ALAMA AND REINHARD KAHLE

COMPUTING WITH MATHEMATICAL ARGUMENTS

ABSTRACT

Thanks to developments in the last few decades in mathematical logic and computer science, it has now become possible to formalize non-trivial mathematical proofs in essentially complete detail. We discuss the philosophical problems and prospects for such formalization enterprises. We show how some perennial philosophical topics and problems in epistemology, philosophy of science, and philosophy of mathematics can be seen in the practice of formalizing mathematical proofs.

Keywords. Epistemic justification, mathematics, formal proof, inferentialism, philosophy of mathematics

1. INTERACTIVELY FORMALIZING MATHEMATICAL ARGUMENTS

Some of the earliest research in artificial intelligence and computer science was on the formalization of mathematical arguments, specifically, the task of using computers to search autonomously for formal proofs of mathematical claims, for example, by H. Wang (Wang 1960). Wang's groundbreaking research led to automatically found proofs of many theorems of *Principia Mathematica*. Commenting on this work in 1960, Wang envisions the birth of a new branch of logic:

> The time is ripe for a new branch of applied logic which may be called "inferential" analysis, which treats proofs as numerical analysis does calculations. This discipline seems capable, in the not too remote future, of leading to machine proofs of difficult new theorems. An easier preparatory task is to use machines to formalize proofs of known theorems.

Wang distinguishes between the automated search for genuinely new mathematical results from the formalization of known theorems. The former is known as *proof search* and the second as *proof check*. Proof search suffers from well-known complexity and undecidability problems. On the other hand, although it might be very hard to find a proof, to *check* a proof for correctness is, in general, not as complex. In this paper we are interested in this second practice, the formalization and verification of known results. Such work, we urge, provides a fascinating glimpse into (a crystallized form of) mathematical practice, and offers the philosopher of science or mathematics a variety of problems and results. We

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_2,

begin with a background of the relevant technology (Section 2) and then delve into three problems raised by the contemporary practice of formalizing mathematical proofs. Section 3 has two parts. Section 3.1 sketches a problem of indeterminacy of content, stemming from the inferentialism to which proof-checking binds us. Section 3.2 discusses a problem of epistemic regress: how do we check the proof checker? Since we are either unable or unwilling to provide proofs in utterly complete logical detail, we want (or need) to leave some inferences to the computer in the sense that it autonomously searches for a deduction from the assumptions in play to the desired conclusion. But we generally want our proofs to have explanatory value, so we often do not want to completely delegate to the computer the task of filling in gaps, because it can often produce inscrutable (though logically correct) proofs or correctly declare, without proof, that a certain non-trivial inference is correct. Which inferences, then, are obvious (safely left to a computer), and which non-obvious? Section 4 takes up this problem in detail. Our aim is that the reader with a sympathy towards or interest in logic and mathematics, with philosophical sensitivities, will be interested to find some tantalizing philosophical problems awaiting him in this field of contemporary computer science.

2. Proof-checking technology

Mathematical practice raises a number of problems for those interested in argumentation. One important characteristic of mathematical argumentation in its in-principle formalizability. Every mathematical proof, it is said, could be formalized in such a way that, starting from, say, axioms of set theory, one can develop the required background concepts and any needed intermediate lemmas all the way to the target theorem.

We have known since at least Frege about in-principle formalizability of mathematical proofs, but it is safe to say that mathematicians do not in fact carry out their arguments formally, in the logician's sense of the term. An obvious rejoinder to the in-principle formalizability of mathematical arguments is that formalized arguments are simply far too large; only for logically trivial propositions can we give a surveyable, practical formalization. If one has some experience with any of the major proof formalisms now available (Hilbert-style proofs, sequent calculi, or natural deduction in various forms) is is apparent that no one would want to *actually* go through with the details of formalizing, say, the proof that there are infinitely many primes starting from, say, the axioms of set theory, working entirely in, say, a Fitch-style natural deduction formalism. Whatever the merits of formalization, it seems that we need to rest content with in-principle formalizability.

In his *Proofs and Refutations* (1976), I. Lakatos puts the problem thus:

> According to formalists, mathematics is identical with formalized mathematics. But what can one *discover* in a formalized theory? Two sorts of things. *First*, one can discover the solution to problems which a suitably programmed Turing machine could solve in a finite time (such as: is a certain alleged proof a proof or not?). No mathematician is interested

in following out the dreary mechanical 'method' prescribed by such decision procedures. *Secondly*, one can discover the solutions to problems (such as: is a certain formula in a non-decidable theory a theorem or not?), where one can be guided by only by the 'method' of 'unregimented insight and good fortune'.

Developments in mathematical logic and computer science in last 25 years or so have, however, helped to make in-principle formalizability a reality. In-practice formalizability is achieved not by skirting the barriers of complexity (some proofs, if written out completely formally, are shockingly large (Orevkov 1993 and Boolos 1984)) or undecidability (validity of arbitrary first-order formulas is not computable (Cutland 1980)), but by designing suitable proof languages to assist in the construction of formal proofs. The idea is that we specify not all, but some steps of a proof, and leave it to a computer to traverse for itself the gaps. Rather than giving a complete proof, we trace or sketch a path through a proof and leave certain logical steps to a computer to fill. This is the principle activity of **interactive theorem proving**. Man and machine cooperate to construct a goal that, in general, neither could accomplish on its own.

A number of interactive theorem provers are thriving. Some of the major ones include Coq[1], Mizar[2], Isabelle[3], HOL4[4], HOL light[5], HOL Zero[6], and Matita[7], among numerous others. These systems take varying approaches toward representing mathematical arguments. Their underlying logics and motivations differ (some are weak, others strong), and the formats in which proofs are written differs as well.

A conventional distinction separates **imperative** from **declarative** proof styles. In the imperative proof style, one proves a theorem by putting forward instructions which, if followed, transform the statement to be proved, which is not immediately acceptable, into other statements that are immediately acceptable. One can do this in a forward (proceeding from the statement to be proved toward acceptable statements) or backward manner (proceeding from acceptable statements toward the thesis to be proved). In the imperative proof style, what one supplies to the interactive theorem prover is not, per se, a proof, but rather a program which says how to construct a proof, without actually supplying it. In the declarative proof style, one does in fact give a proof by writing it down and giving the proof (possibly with gaps in it) to the interactive theorem prover to check.

No matter the style of proof, one can view a formal proof in most interactive theorem provers as a structure that specifies how claims of the argument are justified by various moves. We begin with an initial thesis, and then make inferential

1 http://coq.inria.fr/
2 http://mizar.org
3 http://www.cl.cam.ac.uk/research/hvg/isabelle/
4 http://hol.sourceforge.net/
5 http://www.cl.cam.ac.uk/ jrh13/hol-light/
6 http://proof-technologies.com/holzero.html
7 http://matita.cs.unibo.it/

moves from it, making in turn additional claims. Each step we make transforms the thesis (and possibly introduces new theses) into a different claim. The argument can be said to be successful if all our steps are the result of sound applications of rules of inference and the thesis to be proved at the end of the argument is acceptable.

3. Problems for formal proofs

The practice of formalizing mathematical arguments, although a rather narrow activity, nonetheless illustrates various philosophical problems. In this section we briefly discuss two such problems. One is a problem of indeterminacy of the content of formalized mathematical propositions. The second is a epistemic regress problem. In the next section we discuss a third problem, on the concept of obviousness, in greater detail.

3.1 Inferentialism, indeterminacy of content

One can see the practice of theorem proving as a exemplar of inferentialism. One clearly sees this, in the development of formal theories and proofs, by the need to prove certain claims, given others. The computer acts as an impartial judge of the correctness of one's purported inferences.[8] In developing a formal theory with an interactive or automated theorem prover – developing a formal language, stating definitions, and proving theorems – the meaning of one's statements, in this context, consists *only* in their inferential relationships.

Despite the successes in automated reasoning – faster, more efficient algorithms, new and improved decision procedures – and technical advances that make computers ever more powerful, the technology still keeps us far from "computable inferentialism", by which we understand that a computer can simply decide, quickly, whether statements that we're interested are logical consequences of theories that we're interested in. Practitioners in this kind of computational philosophy know well the battery of problems that arise in this setting:

- *Did we really get the definitions right?*

  It can often happen, in the course of developing a theory formally, that one can get definitions wrong. Of course, in some sense, definitions cannot be

8 We should be clear that no computer can act as a "complete" arbiter of questions like these in the sense that it could correctly answer arbitrary questions of the form: "Is the sentence ϕ a logical consequence of the set Γ of assumptions?" For some notions of logical consequence, such as classical propositional logic, the logical consequence relation is, of course, decidable. The standard notion of first-order logical consequence, though, is, however, undecidable. Some theories expressed in the language of first-order logic are decidable, but most theories of foundational interest, such as Peano Arithmetic or Zermelo-Fraenkel set theory, are themselves undecidable.

wrong. But we can fail to express a definition correctly. Can you spot the error in this definition of prime number in the Mizar language?[9]

```
definition
  let p be Nat;
  attr
    p is prime
  means
    p > 0
    & for n being Nat st n divides p holds
        n = 1 or n = p;
```

- *Are the theorems correctly expressed?*

  This problem is similar to the problem about definitions. Here, the problem is that we might not know what we are proving, in the sense that we may not know the consequences of our definitions. According to the previous definition of prime number, many statements about prime numbers are still valid, and even formally provable, such as the theorem that 2 is a prime number, the theorem that if p and q are distinct primes, then p and q are relatively prime (that is, share no common factor), and even the theorem that every natural number greater than 1 has a prime divisor, which is a lemma toward the proof of the fundamental theorem of arithmetic. It seems that in the formal setting, where the meaning of propositions is their inferential relationships among one another, there are situations where what we are proving is diverging from what we intend to prove. (One would run into trouble with the flawed notion of prime number when it comes time to actually the fundamental theorem of arithmetic because the statement of the theorem is false according to the flawed definition of prime number.)

- *Among multiple candidates for developing a theory/defining a concept/formulating a theorem, which should we choose?*

  Consider, for example, the real number π. One can define it as the ratio of the circumference of a circle to its diameter. This definition implicitly requires that the ratio is the same, for all circles, which is not entirely trivial. Alternatively, one could define π as the real number x in the interval $[0, 4]$ for which $\tan(x/4) = 1$.[10] In this case, one has to of course define the tangent function. What is the tangent function? One could define it in

9 According to this definition, the number 1 is a prime number, which is not correct. We may repair the definition by replacing the constraint that p be positive (`p > 0`) with the condition that p be greater than 1 (`p > 1`).

10 This is in fact how it is defined in the Mizar Mathematical Library, the collection of mathematical knowledge that has been formalized in the Mizar system. See http://mizar.org/version/current/html/sin_*cos.html*.

terms of circles in the real plane. Alternatively, one could define the tangent function using complex analysis (as is done by Mizar) and use the Taylor expansion of the exponential function. In light of these alternatives, what, then is the *real* definition of π? There seems to be no definitive answer, of course. From a certain perspective we can see that the various approaches are "equivalent". We may be able to reconstruct these equivalences, even in a formalized setting. But the problem remains that one must *choose* an initial definition from among the various alternatives. If anything motivates the choice of an initial definition, it is convenience, and not a "true" expression of a concept.

We thus see various forms of indeterminacy in formal mathematics. The source of the indeterminacy seems to be rooted in the fact that, when working with an interactive theorem prover, the meaning of propositions consists, in general, of their inferential roles among each other. We don't know the meaning of our formal assertions until we have drawn out consequences, and even then, it seems that we lack a criterion by which we would know that we have drawn out *enough* consequences that we can rest content that we have a full understanding of the meaning of our definitions and theorems. For there can be derivable statements that express unacceptable propositions, and just as well can there be underivable statements that ought to be derivable.

For lack of space we cannot delve into the various forms of indeterminacy manifested in the practice of formalizing arguments. In the next subsection (3.2), we focus on the issue of regress and trust and then turn, in Section 4 to the problem of what counts as an obvious inference.

3.2 Regress

The principle activity of interactive theorem proving is to formalize a preexisting mathematical argument. At the end of one's work, the result is a formalization of the argument that is completely accepted in all detail by the interactive theorem prover. One can view the theorem prover as a neutral third party that can sign off on the logical correctness of an argument. Does this show that one's argument is *really* correct? In the previous section we discussed an indeterminacy that mitigates the correctness judgment of a theorem prover.

But even if we accept that all axioms and all lemmas and all theorems are correct, there remains a further problem: how do we check the checker? It would seem that we have on our hands a regress. We want to know that our proof is completely correct, but it seems that we cannot accept its correctness without accepting the correctness of the checker.

The judgment of the interactive prover is fallible. We know all too well that most computer programs of any substance suffer from errors (bugs). Interactive theorem provers are no exception. This is as known as it is unavoidable. One guiding design principle among some (but not all) interactive theorem provers the so-called **small kernel** principle, whereby the size of the interactive theorem prover

(as a computer program) is kept as small as possible. Such an approach does not eliminate the possibility of bugs, but it focuses the possible locations of bugs to a small range. The foundations of the HOL light interactive theorem prover, for example, are very small – about 500 lines, compared to millions of lines of code for other substantial computer programs. Any error in HOL light's foundations would be traced to that small kernel. Hales writes (Hales 2008):

> Since the kernel [of HOL light] is so small, it can be checked on many different levels. The code has been written in a friendly programming style for the benefit of a human audience. The source code is available for public scrutiny. Indeed, the code has been studied by eminent logicians. By design, the mathematical system is spartan and clean. The computer code has these same attributes. A powerful type-checking mechanism within the programming language [in which HOL light is programmed] prevents a user from creating a theorem by any means except through this small fixed kernel. Through type-checking, soundness is ensured, even after a large community of users contributes further theorems and computer code.

These features of HOL light by no means guarantee that all theorems proved with it are superlatively correct, but it does give us good reason to believe in their correctness, enough that we need not be distracted by our epistemological worries.

Another way to deal with the regress is to translate the results of one interactive theorem prover into some new context and attempt to re-verify the results there. This has already been done in various ways. One noteworthy endeavor (Urban and Sutcliffe 2008) has been the translation of proofs in the Mizar formalism – which one can view as a kind of multi-sorted first-order logic with some mild extensions – into pure, one-sorted first-order logic. The task is then to verify these translated proofs using tools independent of Mizar. This effort was quite successful (more than 99% of the steps of Mizar proofs can be reverified using an independent automated theorem prover; the missing fraction of unverified steps are due not to an error in the Mizar interactive theorem prover but rather in the translation). As before, such translation and cross-verification successes do not prove that Mizar proofs are wholly correct, but it does give us reason to believe in them, if we had doubts.

4. What counts as "obvious"?

In this section we discuss the problem of delimiting what proofs are acceptable from those that are unacceptable (though logically sound).

An enduring problem of the field of interactive theorem proving is to draw a line between inferences that are accepted without proof from those that require further elaboration. One must balance, on the one hand, the desire for possessing explicit argumentation against, on the other hand, the desire for not wanting to go *too* deep into tedious logical details. If we insist on doing everything ourselves, then we miss an opportunity to take advantage of the power of automated reasoning systems. At the other extreme, if we leave as much as possible to the auto-

mated reasoning system, we may very well be disappointed by its results: strange, "inhuman" patterns of reasoning, excessively long proofs, or surprising outright judgments of validity – "yes, the desired conclusion ϕ is a logical consequence of the set Γ of assumptions currently in play" – that are explanatorily unsatisfactory. So we delegate some reasoning tasks to the interactive theorem prover. Despite their limitations, theorem provers can, at times, find proofs (or disproofs) of surprisingly difficult inferences. Although we might be impressed with the power of the theorem prover, we might wish to reject such cases of "deep" mechanically discovered formal proofs because they undermine the goal of understanding precisely how an informal proof works.

To put the problem another way, what we want out of a formal proof is an explanation for why the theorem of interest follows from our background knowledge. We are happy to accept some inferences without further explanation, such as

$$\{4 \text{ is odd}, 5 \text{ is odd}\} \vdash 4 \text{ is odd} \vee 5 \text{ is odd},$$

but other inferences, such as

$$\mathrm{ZFC} \vdash \text{There exist infinitely many prime numbers},$$

ought to come with an explanation.[11] Which mathematical inferences, then, require explanation, and which do not? If we identify "obviousness" with "not requiring explanation", the problem then is: which inferences are obvious?

Is the concept of obviousness clearly delimited? We may need to parameterize it. What counts as obvious to one person at one time might not count as obvious to the same person at a different time, and it might not be obvious to a different person at the same time. Part of the training in mathematics (and other fields, for that matter) is learning to judge certain things as obvious, even though they don't appear at first sight to be so. A teacher might insist to a student (despite his objections) that a certain theorem is obvious, from a certain perspective, to help acclimatize the student to a certain mathematical field.

The concept of obviousness has various dimensions–logical, rhetorical, epistemological, even social. Despite the vagueness of the concept of obviousness, those who design interactive theorem provers generally need to take a stand on the issue because of the need to delimit those inferences that they will accept from those that they will reject as not sufficiently obvious.

Let us take one important case, the case of the Mizar interactive theorem prover (Grabowski et al. 2010). Mizar is based on classical first-order logic and set theory and uses a natural deduction-style proof language. The body of mathematical knowledge that has been formalized so far in Mizar is quite substantial: it contains more than 50,000 theorems and 10,000 definitions covering essentially an entire undergraduate mathematics curriculum, and more.

11 ZFC is Zermelo-Fraenkel set theory with the axiom of choice.

The notion of "obvious" for Mizar is grounded in a proposal due to M. Davis (Davis 1981 and Rudnicki 1987).[12] Davis offered his proposal in the context of a natural deduction project at Stanford. Natural deduction arguments à la Gentzen, Fitch, or Suppes (as with any serious proof formalism) can be rather tedious. Davis saw that the heart of an informal argument is often obscured or diffused if one adheres strictly to the requirements of the formalism. In formal contexts one wants a rule of inference that would allow one to dispense with certain tedious details. What is wanted is a rule of inference that would allow one to draw a conclusion in one step that, if one were to adhere strictly to the requirements of the formalism, would take many steps (and possibly involve new subproofs). Here is Davis's definition (slightly reformulated):

Definition 1 *A logical formula ϕ is an* ***obvious logical consequence*** *of assumptions Γ if there is a proof of ϕ from Γ that in which each quantified formula of Γ is instantiated at most once.*

According to this definition, to draw an obvious logical inference, it is forbidden to use multiple instances of quantified formulas in Γ. One can elect to choose no quantified formulas in Γ at all, so Davis's concept of obvious inference is a generalization of arbitrary classical propositional logical consequence.[13] If one does choose a quantified formula α in Γ, one then chooses an instance α^* of α by plugging in terms for the outermost universally quantified formulas of α. The task is then to formally derive ϕ from Γ and the instances $\alpha^*_n, \ldots$ using *only* propositional logic.

The notion of obvious logical inference, as defined by Davis, clearly does not characterize how we use the vague term "obvious" in the ordinary discourse. Still, Davis has offered a valuable proposal because of its conceptual simplicity and practical efficiency. In general, if we have some quantified formulas Γ and a conclusion ϕ, deciding which instances of formulas in Γ to choose can quickly lead to a vast number of possibilities, if there are function symbols in the language. Davis's proposal limits the search for instances: *choose (at most) one*. When a conclusion ϕ is not an obvious logical consequence of background assumptions Γ, then the inference is "too complicated" to be checked by a machine, and the human formalizer needs to supply more information.

Davis's concept of obviousness can be parameterized. Let us call an inference of the statement ϕ from premises Γ ***k*-obvious** if ϕ is propositional consequence of Γ and at most k instances of universal formulas in Γ. Davis's concept of obviousness now coincides with the notion of 1-obviousness. The notion of 0-obvious just

12 The actual implementation of the notion of obviousness in Mizar diverges somewhat from the definition that we are about to take up. Nonetheless, the notion we are about to define is the main feature of the actual, implemented notion of obviousness in Mizar.

13 One might object to Davis's proposal at this point because arbitrary classical propositional reasoning is known to be an NP-complete problem. In other words, all tautologies are deemed obvious by Davis's notion. We do not take up the problem of whether this fact conflicts with our ordinary notion of "obvious".

means that no first-order reasoning is involved (formulas starting with universal and existential quantifiers are regarded as unanalyzed atomic formulas). What is the difference between 1-obviousness and 2-obviousness? Consider the following example:[14]

```
reserve X, Y, Z for set;

LemmaOne: X c= Y & Y c= Z implies X c= Z;

LemmaTwo: X c= X \/ Y;

LemmaThree: X c= Y implies X \/ Z c= Y \/ Z;

theorem
X c= Y implies X c= Z \/ Y
proof
 assume X c= Y;
 then A1: Z \/ X c= Z \/ Y by LemmaThree;
 X c= Z \/ X by LemmaTwo;
 hence X c= Z \/ Y by A1, LemmaOne;
end;
```

This is a piece of Mizar proof text with five parts. We will now explain each part.

The first line says that, in what follows, the variables `X`, `Y`, and `Z` are to be understood as universally quantified, and will be sets. The next three statements are background lemmas (assigned the labels `Lemma1`, `Lemma1`, and `Lemma3`) that will later be used in the final proof. Lemma 1 (`LemmaOne`) expresses the transitivity of the subset relation: if $X \subseteq Y$ (`X c= Y`) and $Y \subseteq Z$ (`Y c= Z`), then $X \subseteq Z$ (`X c= Z`). Lemma 2 (`LemmaTwo`) says that X is a subset `c=` of the union $X \cup Y$, written `X \/ Y`, of X with any other set Y. Lemma 3 (`LemmaThree`) says that if X is a subset of Y (`X c= Y`), then the union $X \cup Z$ is a subset of the union $Y \cup Z$ (`X \/ Z c= Y \/ Z`).

Lemmas 1, 2 and 3 will be taken for granted, for the sake of discussion. In other words, the text above is not wholly acceptable to Mizar as written, because, it turns out, all three lemmas are not obvious (in the precise sense of the term) to Mizar and therefore require proof.

Our interest is the final result (the `theorem`) of the Mizar text fragment, which expresses the simple result that if X is a subset of Y (`X c= Y`), then X is a subset of the union $Z \cup Y$ for any set Z (`X c= Z \/ Y`). Unlike the case for the three lemmas, a proof of this fact is provided. Let us proceed through it.

14 I thank Artur Korniłowicz for this example, which comes from the Mizar Mathematical Library. See http://mizar.org/version/current/html/xboole_1.html.

We are carrying out a forward reasoning natural deduction-style proof of an implication (`X c= Y implies X c= Z \/ Y` with antecedent `X c= Y` and consequent `X c= Z \/ Y`). The first step assumes the antecedent (`assume X c= Y`). From this assumption we have, by Lemma 3, the inclusion `Z \/ X c= Z \/ Y` (Mizar is also implicitly using the commutativity of the binary union operation in this inference – Lemma 3 has `Z` on the right-hand side of the union, but in the conclusion just drawn it appears on the left-hand side). The notation `A1:` is assigning a label to the statement just concluded; we will use the formula later by referring to its label. The third step in the argument simply applies Lemma 2; we have from it that $X \subseteq Y \cup X$. We do not use the hypothesis of the theorem nor the previously concluded statement to infer the result; this follows from Lemma 2 alone. The final step of the proof (followed by `hence`) is the desired conclusion: we have that $X \subset Z \cup Y$ from

- $X \subseteq Z \cup X$ (the previous line)
- the formula labeled `A1`, and
- the formula labeled `Lemma1`.

It turns out that this argument is optimal in the sense that no step can be removed. The Mizar proof checker rejects a compression of the argument into simply an enumeration of all background lemmas employed in it: if one were to try to justify the final theorem by simply declaring that it follows from the lemmas, i.e.,

```
X c= Y implies X c= Z \/ Y
  by Lemma1, Lemma2, Lemma3;
```

(the proof is now taken away), one finds that the Mizar proof checker rejects the inference. Other attempted compressions, such as removing the intermediate statement `A1`,

```
X c= Y implies X c= Z \/ Y
proof
 assume X c= Y;
 hence X c= Z \/ Y by Lemma1, Lemma2, Lemma3;
end;
```

or dropping the application of Lemma 2,

```
X c= Y implies X c= Z \/ Y
proof
 assume X c= Y;
 then A1: Z \/ X c= Z \/ Y by Lemma3;
 hence X c= Z \/ Y by A1, Lemma1, Lemma2;
end;
```

are all rejected. They are rejected essentially because the proposed inferences are not obvious, in the Davis/Mizar sense: the proposed compressions apparently require *multiple* instances of background universal premises, and this is precisely what is ruled out by the Davis/Mizar notion of obvious inference. We need to instantiate all three lemmas, each in its own way.

But instead of using the notion of obvious/1-obvious, we use 2-obvious? That is, what if we permitted the proof checker to pick two universal premises, rather than one? The result is that the above proof can indeed be compressed:

```
X c= Y implies X c= Z \/ Y by Lemma1, Lemma2;
```

No proof (beyond listing two background assumptions) is needed anymore! Our claim is a 2-obvious consequence of Lemmas 1 and 2 alone. We don't even need the help of Lemma 3, which was essential before when we were operating under the constraint that all inferences must be 1-obvious.

This example shows that, by strengthening the notion of "obvious inference", the result is that one can get away, in general, with shorter proofs. The compression achieved when we used 2-obviousness rather than 1-obviousness allowed us to get away without supplying any proof at all, and we could even get away with one fewer background assumption. But is that what we are striving for? Perhaps to the reader the very short 2-obvious is what we are after. But we can imagine further examples, using 3-obviousness, 4-obviousness, etc., where proofs become increasingly compressed and inscrutable. Moreover, checking $(k + 1)$-obvious inferences (or, rather, purported $(k + 1)$-obvious inferences) is more complex than checking k-obvious inferences, so we pay a price (in terms of time and space) if what we are after is greater proof compression. We cannot settle the issue here; it is clearly a design parameter for interactive theorem provers, and it would seem that there is no "right" k such that the notion of k-obviousness is ideal. Mizar's insistence upon 1-obviousness is motivated by the need for fast proof checking. It would seem, moreover, that the notion of 1-obviousness leads to satisfactorily explanatory proofs.

5. Conclusion

Far from a dry, unenlightening activity, formalizing mathematical arguments interactively, we suggest, vividly illustrates a variety of philosophical problems, new and old. These problems are evidently prompted by advances in computer science. As an illustration of a new problem raised by developments in automated reasoning, we chose the Mizar interactive theorem prover and studied how it implements the notion of obvious inference. Mizar is but one of many mature interactive theorem provers now widely available. Each one comes equipped with its own notion of obviousness. We are thus left with a family of notions. In the notes the reader will find further references to other interactive theorem provers and relevant literature. Calculemus!

Acknowledgement: Both authors were supported by the ESF research project *Dialogical Foundations of Semantics* within the ESF Eurocores program *LogICCC* (funded by the Portuguese Science Foundation, FCT LogICCC/0001/2007). The second author was also supported by the FCT-project *Hilbert's Legacy in the Philosophy of Mathematics*, PTDC/FIL-FCI/109991/2009.

References

Boolos, G., 1984, "Don't Eliminate Cut", in: *Journal of Philosophical Logic* 13, 4, pp. 373-378.

Cutland, N., 1980, *Computability: An Introduction to Recursive Function Theory.* Cambridge: Cambridge University Press.

Davis, M., 1981, "Obvious Logical Inferences", in: *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 530-531.

Grabowski, A., Korniłowicz, A., and Naumowicz, A., 2010, "Mizar in a Nutshell", in: *Journal of Formalized Reasoning*, 3, 2, pp. 153-245.

Hales, T. C., 2008, "Formal Proof", in: *Notices of the American Mathematical Society* 55, 11, pp. 1370-1380.

Lakatos, I., 1976, *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge: Cambridge University Press.

Orevkov, V. P., 1993, *Complexity of Proofs and their Transformations in Axiomatic Theories*, Vol. 128 of *Translations of Mathematical Monographs.* Providence, RI: American Mathematical Society. Translated by Alexander Bochman from the original Russian manuscript, translation edited by David Louvish.

Rudnicki, P., 1987, "Obvious Inferences", in: *Journal of Automated Reasoning* 3, 4, pp. 383-393.

Urban, J., and Sutcliffe, G., 2008, "Atp-based Cross-verification of Mizar Proofs: Method, Systems, and First Experiments", in: *Mathematics in Computer Science* 2, 2, pp. 231-251.

Wang, H., 1960, "Toward Mechanical Mathematics", in: *IBM Journal of Research and Development* 4, 1, pp. 2-22.

Jesse Alama
Center for Artificial Intelligence
New University of Lisbon
P-2829-516 Caparica
Portugal
j.alama@fct.unl.pt

Reinhard Kahle
Center for Artificial Intelligence and Departamento de Matemática-FCT
New University of Lisbon
P-2829-516 Caparica
Portugal
kahle@mat.uc.pt

Dennis Dieks

Is There a Unique Physical Entropy? Micro versus Macro

Abstract

Entropy in thermodynamics is an extensive quantity, whereas standard methods in statistical mechanics give rise to a non-extensive expression for the entropy. This discrepancy is often seen as a sign that basic formulas of statistical mechanics should be revised, either on the basis of quantum mechanics or on the basis of general and fundamental considerations about the (in)distinguishability of particles. In this article we argue against this response. We show that both the extensive thermodynamic and the non-extensive statistical entropy are perfectly alright within their own fields of application. Changes in the statistical formulas that remove the discrepancy must be seen as motivated by pragmatic reasons (conventions) rather than as justified by basic arguments about particle statistics.

1. Entropy in statistical physics

The concept of entropy has become common even in everyday language, in which it rather vaguely refers to disorder, loss of "energy", waste and dissipation. Users of the concept generally take it for granted that in the background there is a precise scientific notion, with which one should be able to justify, at least in principle, this informal parlance. It is therefore perhaps surprising to find that even in the exact sciences entropy is a multi-faceted concept. It is perhaps least controversial in probability and information theory, at least as far as its mathematical expression is concerned: $S = -\sum_i p_i \ln p_i$ is the generally accepted formula for the entropy S of a probability distribution $\{p_i\}$. But even in the mathematical fields of probability and information theory the exact significance of entropy, and the role that it can play in, e.g., decision theoretical contexts, remains to some extent controversial. One might hope that this will be different once the use of entropy in physics is considered. After all, in physics one expects that the term "entropy" will correspond to something that is accessible to measurement–and drastic differences of opinion about something that can be measured would be surprising. It is this physical entropy, in statistical physics and in thermodynamics, that we shall be concerned with in this paper.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_3,

For the case of M equiprobable events, $p_i = 1/M$, the formula $S = -\sum_i p_i \ln p_i$ reduces to $S = \ln M$. Essentially, this is the famous formula $S = k \ln W$ that can be traced back to Ludwig Boltzmann's seminal 1877 paper about the relation between the second law of thermodynamics and probability theory (Boltzmann 1877, 2001). The constant k (Boltzmann's constant) is merely introduced in order to fix the unit; and W is the number of microstates corresponding to a given macrostate – it is a number of possibilities like M in the earlier formula. The macrostate is defined by macroscopic quantities like pressure, volume and temperature (in the case of a gas in a container); W is the number of microscopic states, characterized by the positions and velocities of the atoms or molecules in the gas, that each give rise to the same values of these macroscopic quantities and in this sense belong to the same macrostate. Boltzmann's entropy thus is basically the earlier introduced S for the case of a probability distribution that assigns equal probabilities to all microstates belonging to a given macrostate. Since the microstates can be represented as points in the phase space of the physical system, the formula $S = k \ln W$ tells us that the entropy of a macrostate is proportional to the logarithm of the volume in microscopic phase space that corresponds to the macrostate.

A paradigmatic and simple application of $S = k \ln W$ is the case of N classical particles (atoms or molecules), each of which can be in any one of X possible states. In this case we find $W = X^N$, and therefore $S = kN \ln X$.

2. Entropy in thermodynamics

In thermodynamics, physical systems are considered from a purely macroscopic point of view. In the case of a gas in a container one looks at changes in macroscopically measurable quantities when the pressure P, volume V and temperature T are made to vary. An essential result, at the basis of the so-called second law of thermodynamics, is that different ways of going from one macroscopic state A to another macroscopic state B (for example, by either first compressing and then cooling, or doing these things in reversed order) are generally associated with different amounts of exchanged heat ΔQ. The heat content of a physical system is therefore not a quantity fixed by its macroscopic state: it is not a *state function*. However, the quantity $\int_A^B dQ/T$, i.e. the exchanged heat divided by the temperature, integrated along a path from A to B (in the macroscopic state space) that represents a *reversible* process, *is* path-independent. That means that $\int_O dQ/T$ does define a state function (the choice of the fiducial state O defines the zero of this function; different choices of O lead to functions that differ by a constant). It is this macroscopic state function that defines the thermodynamic entropy: $S \equiv \int dQ/T$.

Boltzmann's seminal 1877 idea was that the statistical entropy $S = k \ln W$ (Boltzmann himself used another notation) is the microscopic counterpart of the thermodynamic entropy. Each macroscopic state corresponds to a volume in phase space on the micro level, namely the volume occupied by all those microstates that

give rise to the macrostate in question; and the logarithm of this volume represents (apart from immaterial constants) the thermodynamic entropy of the macrostate.

3. A DISCREPANCY

If the micro and macro entropies stand for one and the same physical quantity, the two entropies should obviously depend in exactly the same way on all variables. As it turns out, however, this necessary requirement is not fulfilled. The macro-entropy is *extensive*: if we scale up a physical system by increasing its particle number, its energy and its volume by a factor λ, its entropy will increase by this same factor λ. In other terms, $S(\lambda N, \lambda V, \lambda E) = \lambda S(N, V, E)$. But the micro-entropy as defined above is not extensive.

To see this, imagine two gas-filled chambers of the same volume, separated by a partition. Both chambers contain equal amounts of the same gas in equilibrium, consisting of the same number N of particles. Both parts have the same total energy, temperature T and pressure. Now the partition is removed. What happens to the entropy?

According to thermodynamics the entropy remains the same, because the macroscopic properties of the gases do not change. Smooth removal of the partition is a reversible process without heat transfer; therefore $S_A = S_B$, with A and B the macrostates before and after the removal, respectively. So the total entropy of the double amount of gas, without the partition, is the same as the combined entropy of the two original volumes, i.e. double the entropy of each of the two halves (in this it has been taken for granted that the entropy of several isolated systems is additive – see van Kampen 1984).

However, from the microscopic point of view, the number of available states per particle doubles when the partition is taken out: each particle now has twice as much phase space available to it as it had before. If the number of available states per particle was X with the partition still in place, it becomes $2X$ after the removal of the partition. This means that the number of microstates goes up, from $W_A = X^{2N}$ to $W_B = (2X)^{2N}$, which corresponds to an entropy difference $S_B - S_A = 2kN \ln 2$.

This discrepancy, known as (a version of) the Gibbs paradox, shows that although the thermodynamic entropy is extensive (it doubles when the amount of gas is doubled), the statistical mechanical entropy is not. If we think that there is one and only one physical entropy, this difference between the two approaches signals a problem that needs to be solved. Only one of the two expressions can be right in this case, and since we can directly measure the thermodynamic entropy, and verify its value, it seems clear that the Boltzmann formula $S = k \ln W$ must be wrong. There are two approaches in the literature that take this line. Both claim that fundamental reasoning, starting from first principles on the microscopic level, will not lead to the expression $S = k \ln W$, but instead to the formula $S = k \ln W/N!$, with N the number of particles. This modification of the expression is sufficient to remove our discrepancy.

Remarkably, the two approaches have diametrically opposite starting points: the first, traditional one claims that the *indistinguishability* of particles of the same kind must be taken into account and that this necessitates the insertion of $1/N!$. The second approach says that the *distinguishability* of classical particles has been neglected.

4. The standard "solution": Indistinguishability of particles of the same kind

The traditional way of responding to the discrepancy between micro and macro entropy is to point out that the particles (atoms or molecules) in the two gas chambers are "identical": since they are all atoms or molecules of the same gas, they all possess the same intrinsic properties (charge, mass, etc.). Therefore, a permutation of two or more of these particles should not lead to a new state: it cannot make a difference whether particle 1 is in state a and particle 2 in state b, or the other way around. Both cases equally represent one particle in a and one particle of the same type in b. If we go along with this, the number of microstates W must be adjusted: for a system of N identical particles it must be a factor $N!$ smaller than what we supposed above. When we now redo the calculation, the removal of the partition between the two chambers changes W from $W_A = X^{2N}/(N!)^2$ to $W_B = (2X)^{2N}/(2N)!$. With the help of Stirling's approximation for the factorial it follows that, in the so-called thermodynamic limit $N \to \infty$, $W_B = W_A$. So the total entropy does not change when the partition is taken out: the resulting double-volume amount of gas has double the entropy of each of the separate chambers. This removes the discrepancy between statistical mechanics and thermodynamics.

According to several authors and textbooks, in the final analysis quantum theory is needed for justifying this solution of the Gibbs paradox (see e.g. Schrödinger 1948, Huang 1963, Wannier 1966, Sommerfeld 1977, Schroeder 2000, Ben-Naim 2007). Indeed, classical particles are always distinguishable by their positions, which are strictly correlated to their individual trajectories. These trajectories, in other words the particles' histories, individuate the particles: if we give the particles names on the basis of their positions at one instant, these names persist through time. So the situation in which particle 1 is in state a at a later time is different from the situation in which 2 is in a. It is therefore not self-evident in classical statistical mechanics that we should divide by $N!$. Identical quantum particles, on the other hand, seem indistinguishable in the required sense from the start, because quantum states of systems of identical particles must either be symmetrical under permutation (bosons) or anti-symmetrical (fermions): exchange of particles leaves the state therefore invariant (apart from a global phase factor) and the multiplicity $N!$ never enters.

If this argument were correct, then the non-extensivity of the Boltzmann entropy would show that classical physics is inconsistent and that the world must be quantum mechanical. But obviously, it is hard to believe that simple considerations

about doubling amounts of gases could produce such fundamental insights. Unsurprisingly therefore, doubts have been expressed concerning the just-mentioned traditional solution of the paradox. For example, some authors have claimed that identical classical particles are also fully indistinguishable, and that this justifies the factor $1/N!$ without any recourse to quantum mechanics (e.g., Hestenes 1970, Fujita 1991, Nagle 2004, Saunders 2006).

In the next section we shall take a closer look at whether the permutation of classical particles does or does not make a difference for the microstate.

5. Permutations of "identical" classical particles

We already observed that classical particles can be named and distinguished by their different histories. A process in which two classical particles of the same kind are interchanged can therefore certainly produce a different microstate. Indeed, imagine a situation in which there is one particle at position x_1 and one particle at position x_2, and in which at a later instant there is again one particle at x_1 and one at x_2; suppose that their respective momenta are the same as before. What has happened in the meantime? There are two possibilities: either the particle that was first at x_1 is later again at x_1 and the particle that was first at x_2 is later again at x_2, or the particles have exchanged their positions. The latter case would clearly be different from the former one: it corresponds to a different physical process. Although it is true that the two final situations cannot be distinguished on the basis of their instantaneous properties, their different histories show that the particle at x_1 in one final situation is not the same as the particle at x_1 in the other final situation.

These remarks seem trivial; so what is behind the denial by some authors that identical classical particles can be distinguished and that permutations give rise to different microstates? One reason is that there is an ambiguity in the meaning of the terms "distinguishable" and "permutation". Consider the following statements: "Two particles are distinguishable if they can always be selectively separated by a filter" (Hestenes 1970); "Two particles are distinguishable if they are first identified as 1 and 2, put into a small box, shaken up, and when removed one can identify which particle was the original number 1" (Nagle 2004). With *these* definitions of distinguishability particles of the same kind are indeed indistinguishable. The concept of "permutation" can be interpreted in a similar way. Consider again the microstate of two particles of the same kind, one at x_1 and another at x_2. If the particle at x_2 were at x_1 instead, and the particle at x_1 were at x_2, with all properties interchanged, there would be no physical differences, neither from an observational point of view nor from the viewpoint of theory. One can therefore certainly maintain that the two situations are only two different descriptions (using different ways of assigning indices) for one and the same physical situation (Fujita 1991).

But this is a different kind of permutation from the physical exchange we considered before. In our first example the particles *moved* from x_1 to x_2 and

vice versa. Trajectories in space-time connected the initial state to the permuted state. By contrast, in the alternative reading of "permutation" just mentioned, the exchange is not a physical process at all. Instead, it is an instantaneous swapping that occurs in our thought; it exchanges nothing but indices and does not need trajectories.

A similar sense of "permutation" is employed by Saunders (Saunders 2006). Consider one particle *a* that follows trajectory 1 and another particle *b* of the same kind that follows trajectory 2. Now imagine the case in which particle *a* followed trajectory 2 and particle *b* followed trajectory 1. This exchange would not make any difference for the physical situation. As before, the states before and after a permutation of this kind are not connected by a physical process. A permutation in this sense swaps a supposedly existing abstract "identity" (formally represented by the particle indices "1" and "2", respectively) that is completely independent of the physical characteristics of the situation.

The upshot of these considerations is that if "permutation" is understood as a physical exchange in which trajectories in space-time connect the initial state to the permuted state, then permutations give rise to physically different possibilities, in the sense of different physical processes. If "permutation" is however understood in a different way, then it may well be true that such permutations are not associated with any physical differences and so do not lead to a new microstate.

Let us now consider which kind of permutations is relevant to statistical mechanics – physical exchanges, with connecting trajectories, or swapping indices? Which kind of permutations determines the number of microstates W?

Remember our two gas-filled chambers, each containing N identical particles. Before the removal of the partition the number of available states per particle is X. After the partition has been removed, the number of available states has become $2X$. The reason is that after the partition's removal it has become possible for the particles to *move* to the other chamber. The doubling of the number of available microstates thus expresses a physical freedom that did not exist before the partition was taken away: trajectories have become possible from the particles' initial states to states in the other chamber.

In contrast, even with the partition in place we could consider, in thought, the permutation of "particle identities", or indices, from the left and right sides, respectively – but such permutations are never taken into account in the calculation of the number of microstates. Nor do we consider permutations with particles of the same kind outside of the container, obviously. In other words, the relevant kind of permutations are physical exchanges, not the abstract swapping of indices or identities.

To completely justify the answer that accessibility via a real physical process is the determining factor in the calculation of the number of microstates, we would have to go deeper into the foundations of statistical mechanics. Here, we only mention that one important approach in this area is the ergodic theory, in which the probability of a macrostate is argued to be proportional to the associated volume in phase space on the grounds that this volume is proportional to the amount of

time a system will actually dwell in that part of phase space that corresponds to the macrostate in question. Clearly, this idea only makes sense if the microstates in this part of the phase space are actually accessible via physical trajectories: microstates that give rise to the same macrostate but cannot be reached from the initial situation through the evolution of the system are irrelevant for the macrostate's probability – they do not play a role at all.

It is true that the original form of the ergodic hypothesis (according to which all microstates are actually visited in a relatively short time) has proven to be untenable, but this does not impugn the basic idea that accessibility is the criterion for the relevance of microstates. The multiplicities that occur in more modern and more sophisticated approaches to the foundations of statistical mechanics are the same as those of the original ergodic theory.

We can therefore conclude that in classical statistical mechanics the relevant number of microstates is sensitive to the number of ways this macrostate can be reached via physical processes, i.e. different paths in phase space. Given N particles, there are generally $N!$ different ways in which the particles that have been numbered at some initial time can be distributed in a state at a later time. These permutations represent different physical possibilities, corresponding to different physical processes. Dividing by $N!$ is therefore unjustified when we calculate the numbers of microstates that can be realized by classical particles of the same kind[1].

6. An alternative "solution": Distinguishability of particles of the same kind

In a number of recent publications, Swendsen has proposed an alternative line of reasoning that leads to the entropy formula $S = k \ln W/N!$; he claims that this derivation, rather than the standard accounts, captures the essence of Boltzmann's 1877 ideas (e.g., Swendsen 2002, Swendsen 2008, Swendsen 2012). Swendsen's strategy is to calculate the entropy of a system by considering it as a part of a bigger, composite system; and then to look at the probabilities of microstates of this composite system. Boltzmann's 1877 definition is interpreted as saying that the logarithm of this probability distribution is the entropy of the composite system (apart from multiplicative and additive constants).

Let us illustrate Swendsen's approach by combining a system consisting of a gas of volume V_1 and particle number N_1 with a second gas of the same kind, with volume V_2 and particle number N_2. Let us denote the total volume by V: $V = V_1 + V_2$. The total number of particles, $N = N_1 + N_2$ is taken to be constant (the composite system is isolated), whereas both N_1 and N_2 are variables (the two

1 A more detailed discussion should also take into account that the division by $N!$ is without significance anyway as long as N is constant: in this case the only effect of the division is that the entropy is changed by a constant term $\ln N!$, see (Versteegh 2011).

subsystems can exchange particles). The entropies of both systems, 1 and 2, are now determined in the same derivation.

Swendsen starts from the probability of having N_1 particles in subsystem 1 and $N_2 = N - N_1$ particles in subsystem 2, which for a system of distinguishable individual particles is given by the binomial distribution

$$P(N_1, N_2) = \frac{N!}{N_1!N_2!}\left(\frac{V_1}{V}\right)^{N_1}\left(\frac{V_2}{V}\right)^{N_2}. \tag{1}$$

The entropy of the composite system is subsequently taken to be the logarithm of this probability, plus an arbitrary constant (that only changes the zero of the entropy scale):

$$S(N_1, V_1, N_2, V_2) = k \ln \frac{{V_1}^{N_1}}{N_1!} + k \ln \frac{{V_2}^{N_2}}{N_2!}. \tag{2}$$

In Eq. (2) the value of the additive constant has been set to $k \ln V^N/N!$, for reasons of convenience. It is now clear from Eq. (2) that the entropy of the composite system is the sum of two quantities each of which pertains to only one of the two subsystems. This suggests introducing the function

$$S(N, V) = k \ln \frac{V^N}{N!} \tag{3}$$

as a general expression for the entropy of a system of volume V and particle number N. In the limiting situation in which Stirling's approximation for the factorials applies, taking into account that in thermodynamical equilibrium we will have $V_1/N_1 = V_2/N_2$ (this corresponds to the maximum of the probability distribution), we find that

$$k \ln \frac{{V_1}^{N_1}}{N_1!} + k \ln \frac{{V_2}^{N_2}}{N_2!} \simeq k \ln \frac{V^N}{N!}. \tag{4}$$

This leads to a nicely consistent scheme: if we were to apply the just sketched procedure for finding the entropy to the composite system itself, by combining it with a third system, we would find $S(N, V) = k \ln V^N/N!$ for the entropy of the combined system 1+2. As we now see, this entropy is equal to our earlier defined value in Eq. (2) (fixed by adding the freely chosen constant $k \ln V^N/N!$ to the logarithm of the probability). So we obtain a consistent set of extensive entropies by taking Eq. (3) as our defining equation for entropy.

Swendsen claims that in this way the factor $1/N!$ in the formula for the entropy has been demonstrated to be a necessary consequence of the distinguishability of the gas atoms or molecules. He rejects the formula $S = k \ln W$ and maintains that Boltzmann's ideas, when pursued rigorously like in the just described argument, automatically lead to the expression $S = k \ln W/N!$.

This derivation of $S = k \ln W/N!$ is not convincing, however. First, it should be observed that its starting point, taking the entropy as k times the logarithm of the probability in Eq. (1), is not really different from using the standard formula $S =$

$k \ln W$. This is because the probability $P(N_1, N_2)$ is equal to the volume in phase space measuring the number of states with particle numbers N_1 and N_2, divided by the (constant) total number of states. So the logarithm of the probability is, apart from an additive constant, equal to the logarithm of the number of states with N_1 and N_2. Now, for the comparison with thermodynamics it suffices to replace this number of states with the total number of states: in the thermodynamic limit the probability is peaked, to an extreme degree, around the equilibrium value and the number of equilibrium states is for all practical purposes equal to the total number of states – this is explicitly used by Swendsen in his argument (e.g., Swendsen 2012). Therefore, the entropy of the composite system à la Swendsen is, apart from an additive constant, equal to $S = k \ln W$. Now, what Swendsen effectively does is to fix this additive constant as $1/N!$. There is no problem with this, and exactly the same can be done in the standard approach, since N – the total number of particles in the composite system 1+2 – is a constant. The N-dependency of S that is introduced here is introduced by convention, by choosing a different constant in the definition of S for different values of N.

The next step taken in Swendsen's derivation is to require that the entropy of the system 1+2 should have the same value, and the same N-dependency, in the situation in which it is isolated and the situation in which N is a variable (when 1+2 is brought into contact with a system 3) – this is presented as a requirement of consistency. However, this consistency requirement is exactly the condition that the entropy formula should be such that there will be no change in entropy when a partition is removed. So the derivation boils down to showing that by introducing a N-dependent zero in the definition of the entropy, by convention, the entropy of mixing can be eliminated. But this is what we knew all along! We were asking for a fundamental microscopic justification of the division by $N!$, but Swendsen's argument on close inspection only tells us that the division by $N!$ leads to a convenient expression that makes the entropy extensive and avoids the Gibbs paradox. The insertion of $1/N!$ is in this case just a convention.

This verdict should not be taken as a denial of the fact that the distinguishability of particles is responsible for the occurrence of factorials in expressions in which particle numbers are variables, like (1) and (2). These factorials are important in statistical mechanics, for example in predicting what happens in mixing processes. But it was already argued by Ehrenfest and Trkal (1920, 1921; see also van Kampen 1984) that these factorials can be understood within the standard formalism and do not require a change in the formula $S = k \ln W$ for closed systems. Indeed, the dependence of the total entropy in Eq. (2) on N_1 and N_2 is unrelated to how N occurs in this formula (and to the choice of the zero of the total entropy).

7. The difference between the thermodynamic and statistical entropies

Our original problem was the difference in behavior between the thermodynamic and the statistical entropies: upon removal of a partition between two containers

the entropy increases according to statistical mechanics, whereas it remains the same in thermodynamics. From the point of view of statistical mechanics there is really a change, in the sense that the number of accessible microstates W objectively increases. *In principle* we could verify this empirically, by following the paths of individual particles; we could in this way even measure the microscopic entropy of mixing in a laboratory (Dieks 2010). Admittedly, this would require measurements that lead us outside the domain of thermodynamics. But from the statistical mechanics point of view these changes in phase volume and entropy must be deemed completely natural and objective. This already shows that attempts at eliminating these changes on the basis of arguments on the microscopic scale are doomed to failure. Our analysis of two of such attempts in the previous sections has confirmed this.

This leaves us with the discrepancy between the thermodynamic and statistical entropy. But is there really a problem here? Only if we think of entropy as a Platonic concept that should be the same in all cases (compare van Kampen 1984). If we accept that the two entropies are different, the problem evaporates. After all, entropy is defined differently in statistical mechanics than in thermodynamics: in statistical mechanics the fine-grained micro-description is taken into account as a matter of principle, whereas in thermodynamics this same micro-description is excluded from the start. This difference between the statistical mechanical and the thermodynamical approaches by itself already makes it understandable that the values of entropy changes according to statistical mechanics may sometimes be different from those in thermodynamics (see for a discussion of the consequences of this for the second law of thermodynamics: Versteegh 2011).

From a *pragmatic* point of view it is useful, in many circumstances, if the two theories give us the same entropy values. We can achieve this by a "trick", namely by introducing a new entropy definition in statistical mechanics: replace $S = k \ln W$ by $S = k \ln(W/N!)$. For systems in which N is constant this makes no difference for any empirical predictions: it only adds a constant (though N-dependent!) number to the entropy value. For situations in which N is made to change, this new definition leads to the disappearance of the entropy of mixing and extensivity of the statistical entropy. In this way we obtain agreement with thermodynamics. But it is important to realize that this "reduced entropy" (as it is called by Cheng 2009) has no microscopic foundation; rather, it may be interpreted as the result of a pragmatic decision to erase microscopic distinctions because we are not interested in them in thermodynamics. The division by $N!$ is therefore a convention, motivated by the desire to reproduce thermodynamical results, even though the conceptual framework of thermodynamics is basically different from that of statistical mechanics. The occurrence of $1/N!$ does not necessarily flow from the nature of basic properties of particles, and attempts to prove otherwise are based on a misconception. (Nor should we think that quantum mechanics makes an essential difference here: identical quantum particles can behave just as classical particles in certain circumstances, which again gives rise to the Gibbs paradox; see Dieks and Lubberdink 2011, Versteegh 2011.)

So the solution to our problem is simply to admit that there is a difference between the thermodynamic and the statistical entropy: the thermodynamic entropy is extensive, the statistical entropy is not. Given the different pictures of physical processes painted by thermodynamics and statistical mechanics, respectively, this difference is only natural.

References

Ben-Naim, A., 2007, "On the So-called Gibbs Paradox, and on the Real Paradox", in: *Entropy* 9, pp. 132-136.

Boltzmann, L., 2001, "Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung resp. den Sätzen über das Wärmegleichgewicht", in: *Wissenschaftliche Abhandlungen*, Volume II, pp. 164-224. Providence: AMS Chelsea Publishing.

Cheng, C.-H., 2009,"Thermodynamics of the System of Distinguishable Particles", in: *Entropy* 11, pp. 326-333.

Dieks, D., 2010, "The Gibbs Paradox Revisited", in: *Explanation, Prediction and Confirmation*, edited by D. Dieks et al., pp. 367-377. New York: Springer.

Dieks, D., and Lubberdink, A., 2011, "How Classical Particles Emerge from the Quantum World", in: *Foundations of Physics* 41, pp. 1041-1064.

Ehrenfest, P., and Trkal, V., 1920, "Afleiding van het dissociatie-evenwicht uit de theorie der quanta en een daarop gebaseerde berekening van de chemische constanten", in: *Verslagen der Koninklijke Akademie van Wetenschappen, Amsterdam* 28, pp. 906-929; "Ableitung des Dissoziationsgleichgewichtes aus der Quantentheorie und darauf beruhende Berechnung der chemischen Konstanten", in: *Annalen der Physik* 65, 1921, pp. 609-628.

Fujita, S., 1991, "On the Indistinguishability of Classical Particles", in: *Foundations of Physics* 21, pp. 439-457.

Hestenes, D., 1970, "Entropy and Indistinguishability", in: *American Journal of Physics* 38, pp. 840-845.

Huang, K., 1963, *Statistical Mechanics*. New York: Wiley.

Nagle, J. F., 2004, "Regarding the Entropy of Distinguishable Particles", in: *Journal of Statistical Physics* 117, pp. 1047-1062.

Saunders, S., 2006, "On the Explanation for Quantum Statistics", in: *Studies in the History and Philosophy of Modern Physics* 37, pp. 192-211.

Schrödinger, E., 1948, *Statistical Thermodynamics*. Cambridge: Cambridge University Press.

Schroeder, D. V., 2000, *An Introduction to Thermal Physics*. San Francisco: Addison Wesley Longman.

Sommerfeld, A., 1977, *Thermodynamik und Statistik*. Thun: Deutsch.

Swendsen, R. H., 2002, "Statistical Mechanics of Classical Systems with Distinguishable Particles", in: *Journal of Statistical Physics* 107, pp. 1143-1166.

Swendsen, R. H., 2008, "Gibbs' Paradox and the Definition of Entropy", in: *Entropy* 10, pp. 15-18.

Swendsen, R. H., 2012, "Choosing a Definition of Entropy that Works", in: *Foundations of Physics* 42, pp. 582-593.

van Kampen, N. G., 1984, "The Gibbs Paradox", in: W. E. Parry (Ed.), *Essays in Theoretical Physics*. Oxford: Pergamon Press, pp. 303-312.

Versteegh, M. A. M. and Dieks, D., 2011, "The Gibbs Paradox and the Distinguishability of Identical Particles", in: *American Journal of Physics* 79, pp. 741-746.

Wannier, G. H., 1966, *Statistical Physics*. New York: Wiley.

Institute for History and Foundations of Science
Utrecht University
P.O. Box 80.010
3508 TA, Utrecht
The Netherlands
d.dieks@uu.nl

Luciano Floridi

A Defence of the Principle of Information Closure against the Sceptical Objection

Abstract

The topic of this paper may be introduced by fast zooming in and out of the philosophy of information. In recent years, philosophical interest in the nature of information has been increasing steadily. This has led to a focus on semantic information, and then on the logic of being informed, which has attracted analyses concentrating both on the *statal* sense in which *S* holds the information that *p* (this is what I mean by logic of being informed in the rest of this article) and on the *actional* sense in which *S* becomes informed that *p*. One of the consequences of the logic debate has been a renewed epistemological interest in the principle of information closure (henceforth PIC), which finally has motivated a revival of a sceptical objection against its tenability first made popular by Dretske. This is the topic of the paper, in which I seek to defend PIC against the sceptical objection. If I am successful, this means – and we are now zooming out – that the plausibility of PIC is not undermined by the sceptical objection, and therefore that a major epistemological argument against the formalization of the logic of being informed based on the axiom of distribution in modal logic is removed. But since the axiom of distribution discriminates between normal and non-normal modal logics, this means that a potentially good reason to look for a formalization of the logic of being informed among the non-normal modal logics, which reject the axiom, is also removed. And this in turn means that a formalization of the logic of being informed in terms of the normal modal logic **B** (also known as **KTB**) is still very plausible, at least insofar as this specific obstacle is concerned. In short, I shall argue that the sceptical objection against PIC fails, so it is not a good reason to abandon the normal modal logic **B** as a good formalization of the logic of being informed.

1. Introduction

The topic of this article may be introduced by fast zooming in and out of the philosophy of information.[1] In recent years, philosophical interest in the nature

1 See (Floridi 2011b).

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_4, © Springer Science+Business Media Dordrecht 2013

of information has been increasing steadily.[2] This has led to a focus on semantic information,[3] and then on the logic of being informed,[4] which has attracted analyses concentrating both on the *statal*[5] sense in which *S* holds the information that *p* (this is what I mean by "logic of being informed" in the rest of this article) and on the *actional* sense in which *S* becomes informed that *p*. One of the consequences of the logic debate has been a renewed epistemological interest in the *principle of information closure* (henceforth PIC), which finally has motivated a revival of a sceptical objection against its tenability. Dretske and Nozick (Dretske 1981, 1999, 2006; Nozick 1981) found the objection convincing and their support made it popular. The topic of this article is not a commentary on Dretske's position and the debate that it has generated,[6] but rather a defence of PIC against the sceptical (or, rather, scepticism-based) objection. If I am successful, this means – and we are now zooming out – that the plausibility of PIC is not undermined by the sceptical objection. But since PIC is logically equivalent to the axiom of distribution, this is equivalent to showing that a major epistemological argument against the formalization of the logic of being informed, based on the axiom of distribution in modal logic, is removed. And since the axiom of distribution discriminates between normal and non-normal modal logics, this means that a potentially good reason to look for a formalization of the logic of being informed among the non-normal modal logics,[7] which reject the axiom, is also removed. And this finally means that a formalization of the logic of being informed, in terms of the normal modal logic **B**, is still plausible, at least insofar as this specific obstacle is concerned. In short, I shall argue that the sceptical objection against PIC fails, so the sceptical objection is not a good reason to abandon the normal modal logic **B** as a good formalization of the logic of being informed.

2 For an early overviews see (Floridi 2004).

3 At least since (Dretske 1981), see now (Dretske 1999). For an introduction see (Floridi 2011a).

4 See (Floridi 2006), revised as chapter 10 of (Floridi 2011b).

5 The *statal* condition of being informed is that enjoyed by *S* once *S* has acquired the information (actional state of being informed) that *p*. It is the sense in which a witness, for example, is informed (holds the information) that the suspect was with her at the time when the crime was committed. The distinction is standard among grammarians, who speak of passive verbal forms or states as "statal" (e.g. "the door *was shut* (state) when I last checked it") or "actional" (e.g. "but I don't know when the door *was shut* (act)").

6 On the debate see (White 1991), (Jäger 2004), (Baumann 2006), (Luper 2006), (Shackel 2006), (Dretske 2006). At the time of writing, the most recent contribution is (Adams et al.), which defends Dretske's position. In two recent articles, Genia Schoenbaumsfeld (Schoenbaumsfeld submitted-a, submitted-b) has defended the principle of epistemic closure from a Wittgensteinian perspective that converges with some of the conclusions reached in the following pages. I am grateful to her for sharing her research.

7 The analysis of the logic of being informed in terms of a non-normal modal logic is developed by (Allo 2011).

The paper has the following structure. In Section 2, I formulate PIC against the background provided by the principle of epistemic closure (PEC). There I argue that a satisfactory formulation of PIC is in terms of the *straight principle* of information closure. In Section 3, I formulate the sceptical objection against PIC. In a nutshell, this is a *modus tollens* that holds that PIC is too good to be true: if PIC were acceptable, it would work as a refutation of radical scepticism, yet this violates a more general and widely accepted principle, according to which no amount of factual information can actually answer sceptical questions, so PIC must be rejected. In Section 4, I show that, although the argument is convincing, it mis-allocates the blame: it is not PIC that needs to be abandoned, but the assumption that one might be allowed to start with an uncontroversial piece of factual information, which then provides the input for the correct application of PIC, thus leading to the sceptical refutation. It follows that the sceptical objection does not undermine the tenability of PIC. There might be other good reasons to challenge information closure, but the "too good to be true" argument is not one of them. In Section 5, I consider a potential counter-argument, based on a different formulation of PIC in the context of empirical information processing and show that this too is ineffectual. In the conclusion, I indicate how the acceptance or rejection of PIC determines the choice of normal or non-normal modal logics that best model epistemic and information logics and remind the reader that the removal of the sceptical argument leaves open the choice of a normal modal logic.

2. The formulation of the Principle of Information Closure

Formulating the principle of closure in informational terms is not as straightforward as it might seem. This because PIC is often assumed, at least implicitly, to be a simplified version of the principle of epistemic closure (PEC), and there is quite a large variety of alternative formulations of the latter, each presenting some interesting if subtle mutations.[8] Luckily, the informational translation makes our task less daunting because information is a more impoverished concept than that of knowledge and the ensuing minimalism does help to unclutter our conceptual space. Let us see how.

Initially, it might seem that the best way to formulate PIC would be to use the formulation of PEC under *known* entailment as a template, namely:

K If, while knowing that p, S believes that q because S knows that p entails q, then S knows that q.

K looks like a good starting point because it includes, as an explicit requirement, the fact that S holds (epistemically, doxastically or, in our case, informationally)

8 The interested reader is referred to the excellent review in (Luper 2010). In this article I use K and SP in the way in which they are used in the epistemological literature rather than in modal logic one (see below).

not only that p but also that p entails q. As we shall see presently, this is an advantage, because it enables us to avoid a whole set of distracting issues, based on the contingent or idiosyncratic unavailability of the entailment to a particular S. The fact that Peter might fail to hold the information that Paris is in Europe, while holding the information that Paris is in France, because Peter misses the information that France is in Europe and therefore fails to hold that if Paris is in France then Paris is in Europe, might be relevant in other contexts, e.g. to check how well informed Peter is about European geography, but not here. As it will become clearer in the next two sections, the argument using the sceptical objection attacks PIC not because people have informational or cognitive limits – of course we all do, since we may be distracted, lack a crucial piece of information, be incapable to see what follows from the information that we do hold, run out of time to perform the required logical steps, etc. – but because, if we concede information about both premises, we seem to be able to refute the sceptic, and this, for reasons to be discussed, is alleged to be unacceptable.

The good news is therefore that the requirement of known entailment is a positive feature in K. The bad news is that, despite this, the informational translation of K does not work. Suppose we simplify our task and avoid any reference to beliefs or knowledge. The rationale for this is that we are seeking to formulate a principle of information closure with a broader basis of applicability: it should work for human and artificial agents – including computers that may be able to hold information physically – and hybrid agents, like banks or online services, which might hold information in their files, or in the memories of their employees. Neither artificial nor hybrid agents can be said to *believe* or *know* that p non-metaphorically, for they lack the required mental states or propositional attitudes. In this case, K becomes the principle of known information closure:

PKIC If, while holding the information that p, S holds the information that q because S holds the information that p entails q, then S holds the information that q.

Clearly, PKIC will not do, for it just trivialises the principle into a verbose repetition. If S holds the information that q then S holds the information that q: uncontroversial but also useless. Although it would be interesting to investigate why the informational translation deprives K of its conceptual value, this would go well beyond the scope of this article, so let us not get side-tracked. More constructively, let us keep the known entailment clause in K, which we have seen to be a valuable feature, and use it to modify another version of PEC, known as the *straight principle* of epistemic closure. This states that:

SP If S knows that p, and p entails q, then S knows that q.

The modification, translated in informational terms, gives us:

SPIC If S holds the information that p, and S holds the information that p entails q, then S holds the information that q.

SPIC treats p entails q as another piece of information held by S, as required by the known entailment feature. This avoids contingent or idiosyncratic distractions, as we have seen above in the "French" example with Peter.

Following (Floridi 2006), we obtain what may be called the canonical principle of information closure:

PIC $(Ip \wedge I(p \rightarrow q)) \rightarrow Iq$

PIC is not trivial, or at least not in the sense in which PKIC above is. It also appears to deliver exactly what we need in order to analyse the sceptical objection informationally.

The last step concerns how we *handle* the entailment with the wider scope occurring in PIC. Mind, I do not say *interpret* it, for this is another matter. In the rest of our analysis, I suggest we simplify our task by following the common assumption according to which both entailments are interpreted in terms of material implication. It is the main entailment in PIC that can be handled in several ways. I shall mention two first, for they provide a good introduction to a third one that seems preferable for our current purpose.

A modest proposal is to handle the entailment in terms of *feasibility*. *S could* obtain the information that q, if only S cares enough to extract it from the information that p and the information that p entails q, both of which are already in S' possession. Consider the following example. The bank holds the information that Peter, its chairman, is overpaid. As a matter of fact, the bank also holds the information (endorses the entailment) that, if its chairman is overpaid, then he does not qualify for an annual bonus. So the bank can (but might not) do something with the entailment. Peter might keep receiving his annual bonus for as long as the bank fails to use or indeed decides to disregard the information at its disposal to generate the information that Peter no longer qualifies and then act on it.

A slightly more ambitious proposal, which has its roots in work done by (Hintikka 1962), is to handle the entailment *normatively*: *S should* obtain the information that q. In our example, the bank should reach the conclusion that Peter no longer qualifies for an annual bonus; if it does not, that is a mistake, for which someone (e.g., an employer) or something (e.g., a department) may be reprimanded.

A further alternative, more interesting because it bypasses the limits of the previous two, is to handle the entailment as part of a *sufficient procedure for information extraction* (data mining): in order to obtain the information that q, it is sufficient for S to hold the information that p entails q and the information that p. This third option captures the view that PIC works like an algorithm, with a rule, $I(p \rightarrow q)$, an input Ip and an output Iq. It also leaves unspecified whether S will, can or even should extract q. One way for the bank to obtain the information that Peter does not qualify for an annual bonus is to hold the information that, if he is overpaid, then he does not qualify for an annual bonus, and the information that Peter is overpaid. Handling the entailment as part of a sufficient procedure for

information extraction means qualifying the information that *q* as obtainable independently of further experience, evidence, or input, that is, it means showing that *q* is obtainable without overstepping the boundaries of the available information base. This is just another way of saying that the information in question is obtainable *a priori*.

We now have a satisfactory formulation and interpretation of the principle of information closure. Let us look at the sceptical objection.

3. The sceptical objection

The sceptical objection against PIC has been formulated and debated in several papers. Essentially, it is a *modus tollens*, which requires three steps. The first two are very simple. They consist in providing an interpretation of the information that *p* and of the information that *q* such that *p* entails *q*. The reader is welcome to provide her own version. Here, I shall follow (Kerr and Pritchard forthcoming), and use:

p := *S* is in Edinburgh

q := *S* is not a brain in a vat on Alpha Centauri [henceforth BiVoAC].

e := If *S* is in Edinburgh then *S* is not a brain in a vat on Alpha Centauri [henceforth BiVoAC].

As Kerr and Pritchard remark, referring to Dretske's rejection of PIC:

> […] on Dretske's view I can have an informational basis for believing that I am in Edinburgh but I can have no informational basis for believing that I am not a BIV [brain in a vat] on Alpha Centauri (a skeptical hypothesis which entails that I am not in Edinburgh), even whilst I know that if I am a BIV on Alpha Centauri then I am not in Edinburgh. It is for this reason that Dretske denies epistemic [information] closure.

The third step is the formulation and adoption of a negative thesis:

NT information alone cannot answer a sceptical doubt.

NT seems most plausible. It refers to factual information, and it is a standard assumption in the literature on scepticism, from Sextus Empiricus to Descartes to Wittgenstein. It is explicitly proposed by Dretske himself, shared by Kerr and Pritchard, and I agree with them: sceptical doubts of a Cartesian nature cannot be answered by piling up more or different kinds of factual information. One of the reasons for raising them is precisely because they block such possibility. We would have stopped discussing sceptical questions a long time ago if this were not the case.

We are now ready to formulate the sceptical objection against PIC thus:

i) if PIC, *p* and *e*

ii) then *S* can generate the information that *q a priori*;
iii) but *q* is sufficient for *S* to answer the sceptical doubt (in the example, *S* holds the information that *S* is not a BiVoAC);
iv) and (iii) contradicts NT;
v) but NT seems unquestionable;
vi) so something is wrong with (i)–(iii): in a Cartesian scenario, *S* would simply be unable to discriminate between being in Edinburgh or being a BiVoAC, yet this is exactly what has just happened;
vii) but (iii) is correct;
viii) and the inference from (i) to (ii) is correct;
ix) and *e* in (i) seems innocent;
x) so the troublemaker in (i) is PIC, which needs to be rejected.

It all sounds very convincing, but I am afraid PIC has been framed, and I hope you will agree with me, once I show you by whom.

4. The defence of the principle

Admittedly, PIC looks like the only suspicious character in (i). However, consider more carefully what PIC really achieves, that is, look at *e*. The entailment certainly works, but does it provide any information that can answer the sceptical doubt? Not by itself. For *e* works even if both *p* and *q* are false, of course. This is exactly as it should be, since valid deductions, like *e*, do not generate new information, a scandal (D'Agostino and Floridi 2009) that, for once, it is quite useful to expose. Not only factual information alone cannot answer a sceptical doubt, deductions alone can never answer a sceptical doubt, either. If *e* did generate new information, we would have a bizarre case of synthetic a priori reasoning (recall the handling of the entailment as a sufficient procedure for information extraction), and this seems a straightforward *reductio*. The fact is that the only reason why we take *e* to provide some anti-sceptical, factual information about *S*' actual location in space and time is because we also assume that *p* in *e* is *true*. *Ex hypothesis*, not only *S* is actually in Edinburgh, but *S* holds such information as well. So, if PIC works anti-sceptically, it is because *q* works anti-sceptically, but this is the case because *e* + *p* work anti-sceptically, but this is the case only if *p* is true. Now, *p* is true. Indeed, it should be true, and not just in the chosen example, but in general, or at least for Dretske and anyone else, including myself, who subscribes to the veridicality thesis, according to which *p* qualifies as information only if *p* is true. But then, it is really *p* that works anti-sceptically. All the strength in the antisceptical interpretation of (i)–(iii) comes from the truth of *p* as this is known to *S*, that is, it comes from assuming that *S* is informed that *p*. This becomes obvious once we realise that no shrewd sceptic will ever concede *p* to *S* in the first place, because she knows that, if you concede *p*, then the sceptical challenge is over, as Descartes

correctly argued. Informationally (but also epistemically), it never rains, it pours: you never have just a bit of information, if you have some you *ipso facto* have a lot more. Quine was right about this. Allow a crack in the sceptical dam and the informational flooding will soon be inevitable. In a more epistemological vocabulary, if you know something, you know a lot more than just that something. This is why, in the end, local or circumscribed scepticism is either critical thinking under disguise or must escalate into global scepticism of a classic kind, e.g. Pyrrhonian or Cartesian. The conclusion is that it is really the initial input surreptitiously provided by *p* that is the real troublemaker. PIC is only following orders, as it were. For PIC only exchanges the higher informativeness of a true *p* (where *S* is located, in our example) into the lower informativeness of a true *q* (where *S* is not located, being located where he is). This is like exchanging a twenty pounds banknote into many one-dollar bills. It might look like you are richer, but of course you are just a bit poorer, in the real life analogy because of the exchange rate and the commission charged, and in the sceptical objection because you moved from a positive statement (where you actually are located) to a negative one (one of the infinite number of places where you are not, including places dear to the sceptic like vats in Alpha Centauri). If you do not want the effects of *q*, – if you think that it is rather suspicious to end up with so many dollars coming out of nowhere – do not blame PIC, just never concede *p* in the first place – do not give away the initial British pounds to begin with, using the cash analogy.

It follows that the informational answer to the sceptical doubt, which we agreed was an impossibility, is provided not by *q*, but by *p*, and this disposes of the objection that PIC is untenable because factual information can never provide an answer to sceptical doubts. It never does because one may never be certain that one holds it (one cannot assume to be informed that *p*), not because, if one holds it, it does not.

It might be remarked that all this leaves the last word to the sceptic. I agree, it does, but it does only in this context, and this is harmless. PIC was never meant to provide an antisceptical argument in the first place. It was the alleged accusation that it did in a mistaken way that was the problem. So what happens next? If being in Edinburgh means that I may not be sure that I am there, then we are talking about a scenario in which no further empirical information, no matter how far-reaching, complex, sophisticated or strongly supported, will manage to eradicate once and for all such Cartesian doubt. I believe it is this the proper sense in which all the factual information in the world will never meet the sceptical challenge. For factual information is a matter of empirical facts, and sceptical doubts are based on logical possibilities that challenge the reliability of all such facts. So the no reference to empirical facts, or no offer of factual information can cure logically possible doubts. If you are really worried about being a butterfly that is dreaming to be a human being, showing you that you cannot fly will not work. Is this, then, finally a good reason to reject PIC? The answer is again in the negative. PIC was not guilty when we were assuming to have a foot in the door, a piece of

factual information about how the world really is, namely *p*. It is still not guilty now that we are dealing with a web of information items that might turn out to be a complete fabrication. On the contrary, in the former case it is PIC that helps us to squeeze some (admittedly rather useless) further bits of information from *p*. In the latter case, it is still PIC (though of course not only PIC) that makes the coherence of the whole database of our information tight. But if PIC is to be retained in both cases, what needs to be discharged? Either nothing, if we are allowed a foot in the door, because this is already sufficient to defeat the sceptical challenge; or the value of absolute scepticism as a weapon of total information destruction, if all that it can ever mean is that the logically possible is empirically undefeatable. Once made fully explicit and clarified in detail, radical informational scepticism, with its fanciful scenarios of possible worlds, can be proved to be entirely redundant informationally (Floridi 2010), so it can be disregarded as harmless. Wondering whether we might be dreaming, or living in a Matrix, or might be butterflies who think they are humans, or might be characters in a sci-fi simulation created by some future civilization, and so forth, are interesting speculations that may be intellectually stimulating or simply amusing, but that make no significant difference whatsoever to the serious problem of how we acquire, manage, and refine our information about the world when in the world. The endless game of dealing with them can be left to scholastic philosophers dreaming of final refutations.

5. An objection against the defence and a reply

The reader might still be unconvinced. There might be a lingering doubt about the value of PIC. Such doubt may turn into an objection against the previous defence of PIC that can be formulated by adapting (Adams 2011), who, following Dretske, argues that we should reject information closure. Here it is.

As Adams notices, I too reject PIC in cases in which the kind of information *processing* in question is empirical, as when we see or hear that such and such is the case. As I acknowledged in the past:

> Not all "cognitive" relations are distributive. "Knowing", "believing" and "being informed" are, as well as "remembering" and "recalling". This is why Plato is able to argue that a "mnemonic logic", which he seems to base on *K4*, may replace *DL* [Doxastic Logic] as a foundation for *EL* [Epistemic Logic]. However, "seeing" and other experiential relations, for example, are not: if an agent a sees (in a non metaphorical sense) or hears or experiences or perceives that $p \rightarrow q$, it may still be false that, if a sees (hears etc.) *p*, then a also sees (hears etc.) *q*. (Floridi 2006, p. 441.)

Adams would like to see a more uniform approach and argues that I should simply reject PIC in all cases. I resist it, but we might not be at variance. Consider the following case.

In the left pocket of your jacket you hold the information that, if it is Sunday, then the supermarket is closed. Your watch indicates that today is Sunday. Do you hold the information that the supermarket is closed today? The unexciting answer is maybe. Perhaps, as a *matter of fact*, you do not, so Adams (and Dretske with him) is right. You might fail to make the note in the pocket and the date on the watch "click." Nevertheless, I would like to argue that, as a *matter of logic*, you should, that is, in terms of *feasibility*, *normativity* or *sufficient procedure for information extraction* you did have all the information that the supermarket was closed. So much so that you will feel silly when you are in front of its closed doors and realise that, if you had been more careful, you had all the information necessary to save you the trip. You should have known better, as the phrase goes. Now, I take logic to be a prescriptive not a descriptive discipline. From this perspective, PIC seems to be perfectly fine. This means that the logical application of PIC to informational co-variance is correct. Suppose two systems a and b are coupled in such a way that a's being (of type, or in state) F is correlated to b being (of type, or in state) G, so that $F(a)$ carries (for the observer of a) the information that $G(b)$.[9] An application of PIC in this case means that, if $F(a) \rightarrow G(b)$ qualifies as information and so does $F(a)$, then $G(b)$ qualifies as well. For example, if the low-battery indicator (a) flashing (F) indicates that the battery (b) is flat (G) qualifies as information, and if the battery indicator flashing also counts as information, then so does the battery being flat.

Still from the same perspective, one should not jump to the conclusion that PIC is always applicable to any empirical way of handling information. Consider the example above. This time you *read* the following e-mail, sent by the supermarket: "The shop will be closed every Sunday." You also *read* the date on your computer, which correctly indicates that today is Sunday. Have you *read* that the supermarket is closed today? Of course not, as we assume that there were no further messages. Should you have read that it was? Obviously not, for where was the text that you should have read? Should you have inferred that the supermarket was closed today? Surely, for that was the information that could easily be extracted from the two texts that you read. Again, imagine you are in a hurry and you have only two "time tokens", let us say two seconds. And suppose that reading each message takes one token each. Clearly you do not have the time to extract the information that the supermarket is closed. In more abstract terms, the agent may simply lack the resources to extract not just all (since this is trivially true) but even the relevant information that is logically extractable from the available database.

Adams is talking about the performance of actual players, I am talking about the rules of the game. If Adams's thesis is that PIC is at best only a matter of logic and certainly not an empirical fact, I am convinced.

9 Such co-variance principle has been at the core of the philosophy of information at least since its explicit formulation in (Dretske 1981). The version provide here is from (Floridi 2011b, p. 41), which is a slight modification of the version provided by (Barwise and Seligman 1997).

6. Conclusion: Information closure and the logic of being informed

In this article, I have sought to defend the principle of information closure (PIC) against a popular objection, namely that its assumption would lead to an implausible argument that would defeat radical scepticism. I have shown why such an objection is misdirected. The previous debate might seem to be of interest only to epistemologists or philosophers of information, but such impression would be mistaken. The acceptance or rejection of the principle of closure in epistemology or in the philosophy of information has a wider consequence, in terms of the kind of modal systems that then become available to model epistemic and information logics of different strengths. Quite surprisingly for a topic so well discussed and understood, it seems that such consequence has remained implicit so far, and yet, it is very straightforward. Let me explain. The axiom of distribution states that:[10]

AOD $\quad \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$

AOD discriminates between normal modal logics, to which the axiom applies, and non-normal ones, where the axiom does not apply. PIC is simply the counterpart of AOD in the philosophy of information. This is because PIC can be translated as $(\Box\varphi \wedge \Box(\varphi \rightarrow \psi)) \rightarrow \Box\psi)$, and the latter is logically equivalent to $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$, as a reformulation of both in an implication-free form easily shows. Indeed, AOD is the source of the debate on PEC in modal logic. The parallel is enlightening once it is realised that arguments against AOD in terms of logical omniscience have the same conceptual format as scepticism-based arguments against PIC discussed in above: they are both based on a "too good to be true" strategy.

The fact that PIC and AOD are two sides of the same coin means that the acceptance or rejection of PIC determines whether one is going to consider normal or non-normal modal logics as more suitable to capture all the features one wants to include in an epistemic or information logic. There are good reasons for choosing either option, but two points should now be clear. One is a matter of consistency: rejecting PIC means rejecting the option that epistemic or information logics are normal modal logics. Such rejection is perfectly reasonable and (Allo 2011), for example, offers an interesting analysis of a non-normal alternative. However, and this is the second point, the refutation of the "sceptical argument" against PIC means that one obstacle against a normal modal logic analysis of "*S* is informed that *p*" has been removed. And this, in turn, means that the argument in favour of the analysis of information logic in terms of the *normal* modal logic **B** remains unaffected in this respect.

10 See for example (Cocchiarella and Freund 2008; Hughes and Cresswell 1984). The axiom is also and perhaps better known as the K axiom, but such terminology would be confusing in this paper. A less popular name is deductive cogency axiom.

References

Adams, F., 2011, "Information and Knowledge À La Floridi", in P. Allo (Ed.), *Putting Information First: Luciano Floridi and the Philosophy of Information*. Oxford: Wiley-Blackwell, pp. 84-96.

Adams, F., Barker, J., and Figurelli, J. (forthcoming), "Towards Closure on Closure", in: *Synthese*, pp. 1-18.

Allo, P., 2011, "The Logic of 'Being Informed' Revisited and Revised", in: *Philosophical Studies,* 153, 3, pp. 417-434.

Barwise, J. and Seligman, J., 1997, *Information Flow: The Logic of Distributed Systems.* Cambridge: Cambridge University Press.

Baumann, P., 2006, "Information, Closure, and Knowledge: On Jäger's Objection to Dretske", in: *Erkenntnis,* 64, 3, pp. 403-408.

Cocchiarella, N. B. and Freund, M. A., 2008, *Modal Logic: An Introduction to Its Syntax and Semantics.* New York–Oxford: Oxford University Press.

D'Agostino, M. and Floridi, L., 2009, "The Enduring Scandal of Deduction. Is Propositional Logic Really Uninformative?", in: *Synthese,* 167, 2, pp. 271-315.

Dretske, F., 1981, *Knowledge and the Flow of Information.* Oxford: Blackwell.

Dretske, F., 1999, *Knowledge and the Flow of Information.* Stanford, CA: CSLI Publications.

Dretske, F., 2006, "Information and Closure", in: *Erkenntnis,* 64, 3, pp. 409-413.

Floridi, L., 2004, *The Blackwell Guide to the Philosophy of Computing and Information.* Oxford: Blackwell.

Floridi, L., 2006, "The Logic of Being Informed", in: *Logique et Analyse,* 49, 196, pp. 433-460.

Floridi, L., 2010, "Information, Possible Worlds, and the Cooptation of Scepticism", in: *Synthese* 175, pp. 63-88.

Floridi, L., 2011a, "Semantic Conceptions of Information", in E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* Spring Edition.

Floridi, L., 2011b, *The Philosophy of Information.* Oxford: Oxford University Press.

Hintikka, J., 1962, *Knowledge and Belief: An Introduction to the Logic of the Two Notions.* Ithaca: Cornell University Press.

Hughes, G. E. and Cresswell, M. J., 1984, *A Companion to Modal Logic.* London: Methuen.

Jäger, C., 2004, "Skepticism, Information, and Closure: Dretske's Theory of Knowledge", in: *Erkenntnis* 61, 2-3, pp. 187-201.

Kerr, E. T. and Pritchard, D. (forthcoming), "Skepticism and Information", in: H. Demir (Ed.), *Luciano Floridi's Philosophy of Technology.* New York: Springer.

Luper, S., 2006, "Dretske on Knowledge Closure", in: *Australasian Journal of Philosophy* 84, 3, pp. 379-394.

Luper, S., 2010, "The Epistemic Closure Principle", in: E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*.

Nozick, R., 1981, *Philosophical Explanations.* Oxford: Clarendon Press.

Schoenbaumsfeld, G. (submitted-a), "Mcdowellian Neo-Mooreanism?"

Schoenbaumsfeld, G. (submitted-b), "Meaning and Conversational Impropriety in Sceptical Contexts".

Shackel, N., 2006, "Shutting Dretske's Door", in: *Erkenntnis* 64, 3, pp. 393-401.

White, J. L., 1991, "Knowledge and Deductive Closure", *Synthese* 86, 3, pp. 409-423.

Department of Philosophy
University of Hertfordshire
de Havilland Campus
Hatfield, Hertfordshire AL10 9AB
UK
l.floridi@herts.ac.uk

Hector Freytes, Antonio Ledda, Giuseppe Sergioli and Roberto Giuntini

Probabilistic Logics in Quantum Computation

Abstract

The quantum computation process may be summarized as follows: first an initial state of a physical system is provided as the input. Then, it evolves according to the elementary operations (quantum gates) that are performed on it. Finally, the access to the information content of the resulting state is possible via the measurement operation that provides one of the possible results. In this note we describe probabilistic-type semantics for propositional logics designed to describe effective procedure based on measurement processes.

1. Introduction

Probabilistic logics are conceived to represent the fact that a valid argument is one in which it is not possible for the probability-values of all the premises to be high, while the probability-value of the conclusion is not. More generally, the interest of these logics is to study the propagation of probability-values from the premises to the conclusion of a valid argument. If the premises of a valid argument are all certain, then so is the conclusion. *Probability logics* is the name that Adams [1] proposes for the formal study of the transmission (or lack thereof) of probability-values through valid inferences. Clearly, those basic ideas can be generalized. In fact, alternative axiomatizations of probability defined over event structures different from the usual Boolean σ-algebras bring out alternative logics. This is, in fact, the case of quantum probability [21], or the investigation of states over orthostructures [8, 12, 9]. In this note we describe possible probabilistic semantics arising from probability-values of quantum measurement. The paper is structured as follows: in Section 2, we briefly recall some required basic notions in order to make the paper self-contained. Sections 3 and 4 describe the idea of probabilistic-type logics for pure and mixed states, respectively. Finally, Section 5 outlines possible connections between probabilistic semantics arising from quantum measurement probabilities and fuzzy logic.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_5,

2. Preliminary notions

The notion of *state of a physical system* is familiar in classical mechanics, where it is related to the initial conditions (the initial values of position and momentum) which determine the solutions of the equation of motion of the system. For any value of time, the state is represented by a point in the phase space. In the quantum framework, the description of a state is substantially modified.

Before giving the definition of quantum state, we introduce the concept of *maximal quantum test*. Suppose that we want to observe the properties of a quantum system that can possibly take n different values. If the test you devise allows to distinguish among n possibilities, we say that it is a maximal test. A n-outcome measurement of those properties implements a maximal test. A test that gives only partial information is said to be a partial test. If a quantum system is prepared in such a way that one can devise a maximal test yielding with certainty a particular outcome, then we say that the quantum system is in a *pure state*. The pure state of a quantum system is described by a unit vector in a Hilbert space, and it is denoted by $|\varphi\rangle$ in Dirac notation. If the maximal test for a pure state has n possible outcomes, the state is described by a vector $|\varphi\rangle$ in a n-dimensional Hilbert space. Any orthornormal basis represents a realisable maximal test. Suppose that we have a large number of similarly prepared systems, called an ensemble, and we test for the values of different measurable quantities like, e.g., spin etc. In general, we postulate that, for an ensemble in an arbitrary state, it is always possible to devise a test that yields the n outcomes corresponding to an orthonormal basis with definite probabilities. If the system is prepared in a state $|\varphi\rangle$, and a maximal test corresponding to a basis $\{|e_1\rangle, \ldots, |e_n\rangle\}$ is performed, the probability that the outcome will correspond to $|e_i\rangle$ is given by $p_i(|\varphi\rangle) = |\langle e_i|\varphi\rangle|^2$.

The idea of quantum computation was introduced in 1982 by Richard Feynmann and remained primarily of theoretical interest until developments such as, e.g., Shor's factorization algorithm, that triggered a vast domain of research. In a classical computer, the information is encoded in a series of bits, that are manipulated by logical gates, arranged in a suitable sequence to produce the output. Standard quantum computing is based on quantum systems described by finite dimensional Hilbert spaces, specially $\mathbb{C}^2$, the two-dimensional space of the *qbit*. Similarly to the classical computing case, we can introduce and study the behavior of a number of *quantum logical gates* (hereafter quantum gates for short) acting on qbits. Quantum computing can simulate any computation perfomed by a classical system; however, one of the main advantages of quantum computation and quantum algorithms is that they can speed up the processes.

The standard orthonormal basis $\{|0\rangle, |1\rangle\}$ of $\mathbb{C}^2$ (where $|0\rangle = (1, 0)$ and $|1\rangle = (0, 1)$) is called the *logical* (or *computational*) *basis*. Thus, pure states $|\varphi\rangle$ in $\mathbb{C}^2$ are coherent superpositions of the basis vectors with complex coefficients following the Born rule.

Any qbit $|\psi\rangle = c_0|0\rangle + c_1|1\rangle$ may be regarded as a piece of information, where the number $|c_0|^2$ corresponds to the probability-value that the information

described by the basic state $|0\rangle$ is false; while $|c_1|^2$ corresponds to the probability-value that the information described by the basic state $|1\rangle$ is true. The two basis-elements $|0\rangle$ and $|1\rangle$ are usually taken as encoding the classical bit-values 0 and 1, respectively. By these means, a probability value is assigned to a qbit as follows:

Definition 0.0.1 Let $|\psi\rangle = c_0|0\rangle + c_1|1\rangle$ be a qbit. Then its *probability value* is $p(|\psi\rangle) = |c_1|^2$

Generalizing for a positive integer n, *n-qbits* are represented by unit vectors in the 2^n-dimensional complex Hilbert space $\otimes^n\mathbb{C}^2$. A special basis, called the 2^n-*computational basis*, is chosen for $\otimes^n\mathbb{C}^2$. More precisely, it consists of the 2^n orthogonal states $|\iota\rangle$, with $0 \leq \iota \leq 2^n$, where ι is in binary representation, and $|\iota\rangle$ can be seen as the tensor product of the states $|\iota_1\rangle \otimes |\iota_2\rangle \otimes \ldots \otimes |\iota_n\rangle$, with $\iota_j \in \{0, 1\}$. In this case $\otimes^n|1\rangle = (0, 0 \ldots 0, 1)$ is the n-qbit, in the computational basis, encoding the classical bit 1 in $\otimes^n\mathbb{C}^2$.

A n-qbit $|\psi\rangle \in \otimes^n\mathbb{C}^2$ is a superposition of the basis vectors $|\psi\rangle = \sum_{\iota=1}^{2^n} c_\iota|\iota\rangle$, with $\sum_{\iota=1}^{2^n} |c_\iota|^2 = 1$, and the probability assigned to $|\psi\rangle$ is $|c_{0,0\ldots0,1}|^2$.

In the usual representation of quantum computational processes, a quantum circuit is identified with an appropriate composition of *quantum gates*, i.e. unitary operators acting on pure states of a convenient Hilbert space $\otimes^n\mathbb{C}^2$ [20]. Consequently, quantum gates represent time reversible evolutions of pure states of the system.

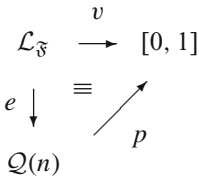
3. Probabilistic-type logic for qbits

Let X be a nonempty set, whose elements are referred as propositional variables, and $\mathfrak{F}$ be a set of connectives each of them with its respective arity. Let $\mathcal{L}_\mathfrak{F}$ be the propositional language from X and $\mathfrak{F}$. A probabilistic logic for qbits may be introduced as a logic $\langle\mathcal{L}_\mathfrak{F}, \models\rangle$, where the propositional variables are interpreted as n-qbits in a given Hilbert space $\otimes^n\mathbb{C}^2$, and the connectives are naturally interpreted as unitary operators acting on pure states in $\otimes^n\mathbb{C}^2$. More precisely, let $\mathcal{Q}(n)$ be the set of n-qbits in $\otimes^n\mathbb{C}^2$ and, if $f \in \mathfrak{F}$, let U_f denote the unitary operator associated to f in $\otimes^n\mathbb{C}^2$.

An interpretation of $\mathcal{L}_\mathfrak{F}$ in $\mathcal{Q}(n)$ is any function $e : \mathcal{L}_\mathfrak{F} \to \mathcal{Q}(n)$ such that, for each $f \in \mathfrak{F}$ with arity k,

$$e(f(x_1, \ldots x_k)) = U_f(e(x_1) \ldots e(x_k)).$$

To define a semantic consequence relation $\models$ from the probability assignment, another step is required: the notion of evaluation. An *evaluation* is any function $v : \mathcal{L}_\mathfrak{F} \to [0, 1]$ such that f factors out as follows:



where p is the probability function in Definition 0.0.1. Hence, the semantic consequence relation $\models$ related to $\mathcal{Q}(n)$ is given by:

$$\alpha \models \beta \text{ iff } (v(\alpha), v(\beta)) \in R$$

with $R \subseteq [0,1]^2$. Since interpretations determine each possible evaluation, for each interpretation e, we denote by e_p the evaluation associated to e.
Hence, a natural extension of the classical logical consequence can be formulated as follows:

$$\alpha \models \beta \text{ iff } (e_p(\alpha) = 1 \text{ implies } e_p(\beta) = 1). \tag{1}$$

A probabilistic logic based on the consequence relation in Condition 1 was developed in [7]. Finally, let us remark that, in [6, 4], the following interesting extension of such a consequence relation was investigated:

$$\alpha \models \beta \text{ iff } e_p(\alpha) \leq e_p(\beta). \tag{2}$$

4. Probabilistic-type logic for mixed states

In general, a quantum system is not in a pure state. This may be caused, for example, by an inefficiency in the preparation procedure of the system, or else because, in practice, systems cannot be completely isolated from the environment, undergoing decoherence of their states. On the other hand, there are interesting processes that cannot be represented as unitary evolutions. A prototypical example of this phenomenon is what happens at the end of a computation process, when a non-unitary operation, a measurement, is applied, and the state becomes a probability distribution over pure states: a *mixed state*.

In view of these facts, several authors [2, 22] paid some attention to a more general model of quantum computational processes, where pure states are replaced by mixed states. This model is known as *quantum computation with mixed states*. Let us briefly describe it.

Let H be a complex Hilbert space. We denote by $\mathcal{L}(H)$ the dual space of linear operators on H. In the framework of quantum computation with mixed states, we regard a quantum state in a Hilbert space H as a *density operator* i.e., an Hermitian operator $\rho \in \mathcal{L}(H)$ that is positive semidefinite ($\rho \geq 0$) and has unit trace ($tr(\rho) = 1$). We indicate by $\mathcal{D}(H)$ the set of all density operators in H.

A *quantum operation* is a linear operator from density operators to density operators such that $\forall \rho \in \mathcal{D}(H) : \mathcal{E}(\rho) = \sum_i A_i \rho A_i^\dagger$, where A_i are operators satisfying $\sum_i A_i^\dagger A_i = I$ and $A_i^\dagger$ is the adjoint of A_i. In the representation of quantum computational processes based on mixed states, a quantum circuit is a circuit whose inputs and outputs are labeled by density operators, and whose gates are labeled by quantum operations. In terms of density operators, an n-qbit $|\psi\rangle \in \otimes^n \mathbb{C}^2$ can be represented as a matrix product $|\psi\rangle\langle\psi|$. Moreover, we can associate to any unitary operator U on a Hilbert space $\otimes^m \mathbb{C}^2$ a quantum operation $\mathcal{O}_\mathcal{U}$, such that, for each $\rho \in \mathcal{D}(H)$, $\mathcal{O}_\mathcal{U}(\rho) = \mathcal{U}\rho\mathcal{U}^\dagger$. Apparently, quantum computation with mixed states generalises the standard model based on qbits and unitary transformations. We would like to stress that the measurement process itself can be also described by a quantum operation, an important fact that strengthnes the choice of quantum operations as representatives of quantum gates. We refer to [2, 20, 22], for more details and motivations about quantum operations. In this powerful model we can naturally extend the logical basis and the notion of probability assignment defined in the qbit case. In fact, we may relate to each vector of the logical basis of $\mathbb{C}^2$ one of the distinguished density operators $P_0 = |0\rangle\langle 0|$ and $P_1 = |1\rangle\langle 1|$, that represent the falsity-property and the truth-property, respectively. The falsity and truth-properties can be generalised to any finite dimension n in the following way:

$$P_0^{(n)} = \frac{1}{tr(I^{n-1} \otimes P_0)} I^{n-1} \otimes P_0 \text{ and } P_1^{(n)} = \frac{1}{tr(I^{n-1} \otimes P_1)} I^{n-1} \otimes P_1,$$

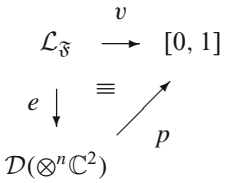
where $n \geq 2$. By the Born rule, the probability to obtain the truth-property $P_1^{(n)}$ for a system in the state ρ is given by the following definition:

Definition 0.0.2 Let $\rho \in \mathcal{D}(\otimes^n \mathbb{C}^2)$.
Then, its *probability value* is $p(\rho) = tr(P_1^{(n)} \rho)$.

Note that, in the particular case in which $\rho = |\psi\rangle\langle\psi|$ where $|\psi\rangle = c_0|0\rangle + c_1|1\rangle$, we obtain that $p(\rho) = |c_1|^2$. Similarly to the case of qbits, we can define a probabilistic logic based on mixed states. Consider the propositional language $\mathcal{L}_\mathfrak{F}$ introduced in Section 3. Here, propositional variables are interpreted as density operators in $\mathcal{D}(\otimes^n \mathbb{C}^2)$, whilst connectives are naturally interpreted as quantum operations acting on $\mathcal{D}(\otimes^n \mathbb{C}^2)$. If f is a connective in $\mathfrak{F}$, we denote by $\mathcal{E}_f$ the quantum operation associated to f in $\mathcal{L}(\mathbb{C}^2)$. An interpretation of $\mathcal{L}_\mathfrak{F}$ in $\mathcal{D}(\otimes^n \mathbb{C}^2)$ is any function $e : \mathcal{L}_\mathfrak{F} \to \mathcal{D}(\otimes^n \mathbb{C}^2)$ such that for each $f \in \mathfrak{F}$ with arity k,

$$e(f(x_1, \ldots x_k)) = \mathcal{E}_f(e(x_1) \ldots e(x_k)).$$

To define a semantic consequence relation $\models$, we also consider a natural adaptation of the notion of evaluation. Accordingly, in this setting an *evaluation* will be any function $v : \mathcal{L}_\mathfrak{F} \to [0, 1]$ that makes our diagram commutative:



where p is the probability function in Definition 0.0.2. Hence, the semantic consequence $\models$ related to $\mathcal{D}(\otimes^n \mathbb{C}^2)$ will be:

$$\alpha \models \beta \quad \text{iff} \quad (v(\alpha), v(\beta)) \in R, \tag{3}$$

with $R \subseteq [0, 1]^2$.

5. Connections with fuzzy logic

Since the Eighties, the interest in many-valued logics enormously increased. In particular, the so called *fuzzy logics*, with their truth values in the real interval $[0, 1]$, emerged as a consequence of the 1965 proposal, by L. Zadeh, of a fuzzy set theory [23]. A fundamental system of fuzzy logic, introduced by P. Hájek in [13], is known as *basic fuzzy logic*. A relevant feature of those logics is the notion of conjunction whose natural interpretation is a real valued function in $[0, 1]$, that goes under the name of continuous t-norm. More precisely, a t-norm is a continuous binary function $\odot : [0, 1]^2 \to [0, 1]$ that satisfies the following conditions:

1. $x \odot 1 = x$;
2. if $x_1 \leq x_2$ and $y_1 \leq y_2$, then $x_1 \odot y_1 \leq x_2 \odot y_2$;
3. $x \odot y = y \odot x$;
4. $x \odot (y \odot z) = (x \odot y) \odot z$.

In [16], K. Menger used the idea of t-norm in the framework of the probabilistic metric spaces. In such spaces, t-norms allow us to generalise the triangle inequality for probability distribution valued metrics. In basic fuzzy logic a genuine relationship between conjunction and implication can be established. In this system the continuity of the t-norm plays an important role. The following are the three basic continuous t-norms:

- $x \odot_P y = x \cdot y$, (Product t-norm);
- $x \odot_Ł y = \max\{x + y - 1, 0\}$, (Łukasiewicz t-norm);
- $x \odot_G y = \min\{x, y\}$, (Gödel t-norm).

These t-norms are remarkably basic, in that each possible continuous t-norm can be obtained as an *adequate combination* of them [15]. Further it is interesting to

notice that these three t-norms always represent irreversible functions. In [11] is introduced a special type of quantum operations called *polynomial quantum operation*. These quantum operations can probabilistically represent any polynomial function g such that:

(1) each coefficent of g lives in $[0, 1]$;

(2) the restriction $g \restriction_{[0,1]^n}$ lives in $[0, 1]$.

Further, in [11] is shown that any continuous function (not necessarily polynomial) that satisfy conditions (1) and (2) can be *approximately* represented by means of a polynomial quantum operation. Not surprisingly, the accuracy of the approximation is arbitrary (the higher is the accuracy of the approximation, the higher is the quantum operation complexity degree). The three t-norms previously introduced are continuous functions that satisfy conditions (1) and (2). Accordingly by the results mentioned above, for each of the three t-norms there exists a polynomial quantum operation that represents it. Further, in case of Product t-norm this representation is *exact* (since Product t-norm is polynomial), while in the other two cases is *approximated*.

The representation of continuous t-norms as quantum operations motivated the investigation of a logical system in the framework of probabilistic-type logic for mixed states.

Let us recall the following definition first. The *standard PMV-algebra* (*standard product multi-valued algebra*) [10, 19] is the algebra

$$[0, 1]_{PMV} = \langle [0, 1], \oplus, \odot_P, \neg, 0, 1 \rangle,$$

where $[0, 1]$ is the real unit segment, $x \oplus y = \min(1, x + y)$, the operation $\odot_P$ is the real product (corresponding to the Product t-norm introduced above), and $\neg x = 1 - x$. A slight weakening of this structure (called *quasi PMV-algebra*) plays a notable role in quantum computing, in that it describes, in a probabilistic way, a relevant quantum gates in the framework of *Poincarè irreversible quantum computational algebras* [5, 7].

As is well known, fuzzy logics (and infinite-valued Łukasiewicz logic in particular) play a relevant role in game theory and theoretical physics as shown in [17, 18], where it is investigated the deep connection between infinite-valued Łukasiewicz logic with Ulam games and $AF-C^*$-algebras. It would be desirable to extend this connection, by means of quasi-PMV algebras, to the investigation of quantum games and to error-correction codes in the context of quantum computation.

References

[1] Adams, E., 1998, *A Primer of Probability Logic*. Stanford: CSLI, Stanford University.

[2] Aharonov, D., Kitaev, A., and Nisan, N., 1997, "Quantum Circuits with Mixed States", in: *Proc. 13th Annual ACM Symp. on Theory of Computation*, pp. 20-30.

[3] Beltrametti, E., Dalla Chiara, M. L., Giuntini, R., Leporini, R., and Sergioli, G., 2012, "Epistemic Quantum Computational Structures in a Hilbert-space Environment", in: *Fundamenta Informaticae* 20, pp. 1-14.

[4] Bou, F., Paoli, F., Ledda, A., Spinks, M., and Giuntini, R., 2010, "The Logic of Quasi MV Algebras", in: *Journal of Logic and Computation* 20, 2, pp. 619-643.

[5] Cattaneo, G., Dalla Chiara, M. L., Giuntini, R., and Leporini, R., 2004, "Quantum Computational Structures", in: *Mathematica Slovaca*, 54, pp. 87-108.

[6] Dalla Chiara, M. L., Giuntini, R., and Greechie, R., 2004, *Reasoning in Quantum Theory*. Dordrecht: Kluwer.

[7] Domenech, G., and Freytes, H., 2006, "Fuzzy Propositional Logic Associated with Quantum Computational Gates", in: *International Journal of Theoretical Physics* 34, pp. 228-261.

[8] Domenech, G., Freytes, H., and de Ronde, C., 2011, "Equational Characterization for Two-valued States in Orthomodular Quantum Systems", in: *Reports on Mathematical Physics* 68, pp. 65-83.

[9] Dvurečenskij, A., and Pulmannová, S., 2000, "New Trends in Quantum Structures", Vol. 516 of *Mathematics and Its Applications*, Dordrecht: Kluwer.

[10] Esteva, F., Godo, L., and Montagna, F., 2001, "The $L\Pi$ and $L\Pi\frac{1}{2}$: Two Complete Fuzzy Systems Joining Łukasiewicz and Product Logic", in: *Archive for Mathematical Logic* 40, pp. 39-67.

[11] Freytes, H., Sergioli, G., and Aricó, A., 2010, "Representing Continuous T-norms in Quantum Computation with Mixed States", in: *Journal of Physics A* 43, pp. 1-15.

[12] Gudder, S., 1979, *Stochastic Methods in Quantum Mechanics*. North Holland–New York: Elsevier.

[13] Hájek, P., 1998, *Metamathematics of Fuzzy Logic*. Dordrecht: Kluwer.

[14] Lawler, E. L., and Sarkissian, I. S., 1995, "An Algorithm for Ulam's Game and its Application to Error Correcting Codes", in: *Information Processing Letters* 56, pp. 89-93.

[15] Ling, C. M., 1965, "Representation of Associative Functions", in: *Publicationes Mathematicae Debrecen*, 12, pp. 189-212.

[16] Menger, K., 1942, "Statistical Metrics", in: *Proceedings of the National Academy of Sciences of the United States of America* 37, pp. 57-60.

[17] Mundici, D., 1986, "Interpretation of AF C^*-algebras in the Łukasiewicz Sentential Calculus", in: *Functional Analysis* 65, pp. 15-63.

[18] Mundici, D., 1993, "Ulam Games, Łukasiewicz Logic and AF C^*-algebras", in: *Fundamenta Informaticae* 18, pp. 151-161.

[19] Mundici, D., and Riecăn, B., "Probability on MV-algebras", in: E. Pap (Ed.), *Handbook of Measure Theory.* Amsterdam: Elsevier, pp. 869-909.

[20] Nielsen, M. A., and Chuang, I. L., 2000, *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press.

[21] Rédei, M., and Summers, S. J., 2007, "Quantum Probability Theory", in: *Studies in the History and Philosophy of Modern Physics* 38, pp. 390-417.

[22] Tarasov, V., 2002, "Quantum Computer with Mixed States and Four-Valued Logic", in: *Journal of Physics A* 35, pp. 5207-5235.

[23] Zadeh, L., 1965, "Fuzzy Sets", in: *Information and Control* 8, pp. 338-353.

Hector Freytes
Instituto Argentino de Matematica (IAM)
Saavedra 15
1083, Buenos Aires
Argentina
hfreytes@gmail.com

Antonio Ledda, Giuseppe Sergioli, Roberto Giuntini
Faculty of Education
University of Cagliari
Via Is Mirrionis, 1
09123, Cagliari
Italy
antonio.ledda@unica.it
giuseppe.sergioli@gmail.com
giuntini@unica.it

ALEXEI GRINBAUM

QUANTUM OBSERVER, INFORMATION THEORY AND KOLMOGOROV COMPLEXITY

ABSTRACT

The theory itself does not tell us which properties are sufficient for a system to count as a quantum mechanical observer. Thus, it remains an open problem to find a suitable language for characterizing observation. We propose an information-theoretic definition of observer, leading to a mathematical criterion of objectivity using the formalism of Kolmogorov complexity. We also suggest an experimental test of the hypothesis that any system, even much smaller than a human being, can be a quantum mechanical observer.

1. INTRODUCTION

A few years after Carlo Rovelli proposed a relational interpretation of quantum mechanics (Rovelli 1996), it received a sharp rebuke from Asher Peres. The issue that Peres addressed was Rovelli's claim to the universality of the quantum mechanical observer. According to Rovelli, all systems should be seen as observers insofar as their degrees of freedom are correlated with the degrees of freedom of some other system. Information contained in such a correlation is the information possessed by the observer about the observed system. Nothing else is needed, not even a limit on the size of systems or the number of their degrees of freedom. This is where Peres objected: "The two electrons in the ground state of the helium atom are correlated, but no one in his right mind would say that each electron 'measures' its partner" (Peres 1986). The controversy is still unresolved: Is the capacity to serve as quantum mechanical observer universal and extends to all systems? Or is it true that only some systems, but not others, can be observers, and if there is a limitation, then what is it precisely?

I will argue that in order to give an answer to this question, we need to revolutionize our idea of physical observation. For this, I'll first briefly review the history of thinking on quantum mechanical observers and then I'll propose a new conceptual toolkit with which to approach this question. This toolkit will involve the notion of information and the Kolmogorov complexity as its quantitative measure.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_6, © Springer Science+Business Media Dordrecht 2013

2. Observer in the interpretations of quantum mechanics

2.1 Observer in the Copenhagen orthodoxy

Bohr's lecture at Como in 1927 was a foundation of what later came to be known as the Copenhagen interpretation of quantum mechanics. Despite being a common reference among physicists, this interpretation has a variety of slightly different formulations. Its main point, however, is clearly stated by Bohr:

> Only with the help of classical ideas is it possible to ascribe an unambiguous meaning to the results of observation.... It lies in the nature of physical observation, that all experience must ultimately be expressed in terms of classical concepts. (Bohr 1934, p. 94)

Two different readings of this statement are possible, divided by what exactly is meant by "classical". The first reading is a straightforward *sine qua non* claim about quantum and classical mechanics:

> It is in principle impossible to formulate the basic concepts of quantum mechanics without using classical mechanics. (Landau and Lifschitz 1977, p. 2)

The second reading is that quantum mechanical experiments can only be described by classical language. Even if classical language later leads us to classical mechanics, it is the *language* – not any form of mechanics – that becomes a crucial ingredient:

> Bohr went on to say that the terms of discussion of the experimental conditions and of the experimental results are *necessarily* those of 'everyday language', suitably 'refined' where necessary, so as to take the form of classical dynamics. It was apparently Bohr's belief that this was the only possible language for the *unambiguous communication* of the results of an experiment. (Bohm 1971, p. 38)

The first reading implies that the world consists of mechanical systems only, whether quantum or classical, and no observer external to physical theory is necessary. Contrary to this, the second reading assumes that the formulation of the problem includes an agent possessing classical language: the experimenter. The latter prepares and measures the quantum system, thereby acting as a quantum mechanical observer.

2.2 London and Bauer

First published in 1932, John von Neumann's magisterial book on quantum mechanics offered what were to become a standard theory of quantum measurement (von Neumann 1932). But von Neumann's musings about the place of the observer during measurement were not entirely satisfactory. The mathematics worked perfectly, however its meaning required further clarification. Writing as early as 1939,

London and Bauer set the tone of the conceptual debate. They noted that quantum mechanics didn't ascribe properties to the quantum system in itself, only in connection to an observer. For London and Bauer such an observer had to be human: "it seems that the result of measurement is intimately linked to the consciousness of the person making it" (London and Bauer 1939, p. 48). The cut between the observer and the observed system introduced by von Neumann and Dirac was pushed to the extreme, leaving all physical systems – even the human eye and the visual nerve – on one side, and only leaving the observer's 'organ' of awareness, namely consciousness, on the other.

If this were true, why would objectivity be possible at all and why have physicists not yet become solipsists? Why do two physicists agree on what constitutes the object of their observation and on its properties? According to London and Bauer, the reason is the existence of something like a "community of scientific consciousness, an agreement on what constitutes the object of the investigation"(London and Bauer 1939, p. 49). The exact meaning of this assertion remained a mystery.

2.3 Wigner

Bohr emphasized that a linguistic faculty is necessary for observers because they must communicate unambiguously. This was further developed by Eugene Wigner. The consciousness of the observer "enters the theory unavoidably and unalterably" and corresponds to an impression produced by the measured system on the observer. The wave function "exists" only in the sense that "the information given by the wave function is communicable":

> The communicability of information means that if someone else looks at time t and tells us whether he saw a flash, we can look at time $t + 1$ and observe a flash with the same probabilities as if we had seen or not seen the flash at time t ourselves. (Wigner 1961)

The observer "tells us" the result of his measurement: like for Bohr, communication for Wigner is therefore linguistic. But do observers actually *have* to communicate or is it enough to require that they simply *could* communicate? On the one hand, Wigner says, "If someone else somehow determines the wave function of a system, he can tell me about it…", which requires a mere possibility of communication but no sending of actual information. On the other hand, he famously analyzes the following 'Wigner's friend' situation:

> It is natural to inquire about the situation if one does not make the observation oneself but lets someone else carry it out. What is the wave function if my friend looked at the place where the flash might show at time t? The answer is that the information available about the *object* cannot be described by a wave function. One could attribute a wave function to the joint system: friend plus object, and this joint system would have a wave function also after the interaction, that is, after my friend has looked. I can then enter into interaction with this joint system by asking my friend whether he saw a flash. … The typical change in the

wave function occurred only when some information (the *yes* or *no* of my friend) entered *my* consciousness. (Wigner 1961)

Although he calls this situation natural, Wigner is the only one among the founding fathers of quantum theory to have addressed it explicitly. Here Wigner's agreement with his friend is clearly possible thanks to the linguistic communication between them, but this communication itself is not a quantum measurement: whatever the situation, Wigner always knows the question he should put to his friend and fully trusts the answer, always *yes* or *no*. Communication from the friend must actually occur before the wave function could be known by Wigner; it is not enough that this communication be merely possible. The question remains open as for the exact mechanism, whether a human convention or a physical given, of the agreement between observers.

Wigner also touches on the question of belief and trust in his discussion of repeatability of experiments in physics. To explore the statistical nature of the predictions of quantum mechanics, it is necessary to be able to produce many quantum systems in the same state; subsequently these systems will be measured. One can never be absolutely sure, as Wigner stipulates, that one has produced the same state of the system. We usually "believe that this is the case" and we are "fully convinced of all this"(Wigner 1976, p. 267), even if we have not tried to establish experimentally the validity of the repeated preparation of the same state. What is at work here is again a convention shared by all physicists. How do they know that repeated preparations produce the same state if they do not measure each and every specimen in order to verify it? The answer is that they have common experience and a convention on what a 'controlled experiment' amounts to, and their respect of this commonly shared and empirically validated rules enables them to postulate the existence of repeated states even in the situations which had never been tested before. This is how physical theory with its laws and a precise methodology arises by way of abstraction ('elevation', as Einstein or Poincaré would say (Friedman 2001, p. 88)) from the physicist's empirical findings and the heuristics of his work.

2.4 Everett

The need to refer to consciousness exists insofar as only consciousness can distinguish a mere physical correlation, e.g. of an external system with the observer's eye, from the information actually available to the observer, i.e. the observer's knowledge on which he can act at future times. Other characteristics are irrelevant: jokingly, London and Bauer tell us that "there is little chance of making a big mistake if one does not know [the observer's] age" (London and Bauer 1939, p. 43). Treating the observer as an informational agent requires that we say precisely what property authorizes different systems possessing information to be treated as observers. In other words, what is the nature of a convention shared by all observers? Brillouin was among the first to believe that information in physics must be defined with the exclusion of all human element (George 1953, p. 360). This

was continued by Hugh Everett (1957), for whom observers are physical systems that possess memory. Memory is defined as "parts... whose states are in correspondence with past experience of the observers". Thus observers do not have to be human: they could be "automatically functioning machines, possessing sensory apparatus and coupled to recording devices".

Everett was the first to explicitly consider the problem of several observers. The "interrelationship between several observers" is an act of communication between them, which Everett treats as establishing a correlation between their memory configurations. He listed several principles to be respected in such settings:

1. When several observers have separately observed the same quantity in the object system and then communicated the results to one another they find that they are in agreement. This agreement persists even when an observer performs his observation after the result has been communicated to him by another observer who has performed the observation.
2. Let one observer perform an observation of a quantity A in the object system, then let a second perform an observation of a quantity B in this object system which does not commute with A, and finally let the first observer repeat his observation of A. Then the memory system of the first observer will *not* in general show the same result for both observations. ...
3. Consider the case when the states of two object systems are correlated, but where the two systems do not interact. Let one observer perform a specified observation on the first system, then let another observer perform an observation on the second system, and finally let the first observer repeat his observation. Then it is found that the first observer always gets the same result both times, and the observation by the second observer has no effect whatsoever on the outcome of the first's observations. (Everett 1957)

As we shall see, the problem of agreement between different observers and the need for memory as a defining characteristics of observation are intimately connected.

3. Information-theoretic definition of observer

3.1 Observer as a system identification algorithm

What characterizes an observer is that it has information about some physical system. This information fully or partially describes the state of the system. The observer then measures the system, obtains further information and updates his description accordingly. Physical processes listed here: the measurement, updating of the information, ascribing a state, happen in many ways depending on the

physical constituency of the observer. The memory of a computer acting as an observer, for instance, is not the same as human memory, and measurement devices vary in their design and functioning. Still one feature unites all observers: that whatever they do, they do it to a *system*. In quantum mechanics, defining an observer goes hand in hand with defining a system under observation. An observer without a system is a meaningless nametag, a system without an observer who measures it is a mathematical abstraction.

Quantum systems aren't like sweets: they don't melt. Take a general thermodynamic system interacting with other systems. Such a system can dissipate, diffuse, or dissolve, and thus stop being a system. If at first a cube of ice gurgling into tepid water is definitely a thermodynamic system, it makes no sense to speak about it being a system after it has dissolved: the degrees of freedom that previously formed the ice cube have been irreparably lost or converted into physically non-equivalent degrees of freedom of liquid water. Quantum systems aren't like this. The state of a quantum system may evolve, but the observer knows how to tell the system he observes from the environment. An electron in a certain spin state remains an electron after measurement even if its state has changed, i.e., it remains a system with a particular set of the degrees of freedom which we call an "electron". Generally speaking, the observer maintains system identity through a sequence of changes in its state. Hence, whatever the physical description of such 'maintaining' may be, and independently of the memory structure of a particular physical observer, first of all every observer is abstractly characterized as a system identification machine. Different observers having different features (clock hands, eyes, optical memory devices, internal cavities, etc.) all share this central feature.

Definition 1. An observer is a system identification algorithm (SIA).

Particular observers can be made of flesh or, perhaps, of silicon. 'Hardware' and 'low-level programming' are different for such observers, yet they all perform the task of system identification. This task can be defined as an algorithm on a universal computer, e.g., the Turing machine: take a tape containing the list of all degrees of freedom, send a Turing machine along this tape so that it puts a mark against the degrees of freedom that belong to the quantum system under consideration. Any concrete SIA may proceed in a very different manner, yet all can be modelled with the help of this abstract construction.

The SIAs with possibly different physical realization share one property that does not depend on the hardware: their algorithmic, or Kolmogorov, complexity. Any SIA can be reconstructed from a binary string of some minimal length (which is a function of this SIA) by a universal machine. As shown by Kolmogorov, this minimal compression length defines the amount of information in the SIA and does not depend (up to a constant) on the realization of the SIA on particular hardware (Kolmogorov 1965).

3.2 Quantum and classical systems

Each quantum system has a certain number of degrees of freedom: independent parameters needed in order to characterize the state of the system. For example, a system with only two states (spin-up and spin-down) has one degree of freedom and can be described by one parameter $\sigma = \pm 1$. If we write these parameters as a binary string, the Kolmogorov complexity of this string is at least the number of the degrees of freedom of the system. Consequently, for any system S and the Kolmogorov complexity of the binary string s representing its parameters

$$K(s) \geq \mathrm{d}_S, \tag{1}$$

where d_S is the number of the degrees of freedom in S. In what follows the notation $K(s)$ and $K(S)$ will be used interchangeably.

When we say that observer X observes quantum system S, it is usually the case that $K(S) \ll K(X)$. In this case the observer will have no trouble keeping track of all the degrees of freedom of the system; in other words, the system will not 'dissolve' or 'melt' in the course of dynamics. However, it is also possible that X identifies a system with $K(S) > K(X)$. For such an observer, the identity of system S cannot be maintained and some degrees of freedom will fall out from the description that X makes of S.

Definition 2. System S is called quantum with respect to observer X if $K(S) < K(X)$, meaning that X will be able to maintain a complete list of all its degrees of freedom. Otherwise S is called classical with respect to X.

Suppose that X observes a quantum system, S, and another observer Y observes both S and X. If $K(Y)$ is greater than both $K(X)$ and $K(S)$, observer Y will identify both systems as quantum systems. In this case Y will typically treat the interaction between X and S as an interaction between two quantum systems. If, however, $K(X)$ and $K(Y)$ are close, $K(X) \gg K(S)$ and $K(Y) \gg K(S)$ but $K(X) \simeq K(Y)$, then Y will see S as a quantum system but the other observer, X, as a classical system. An interaction with a classical system, which we usually call 'observation', is a process of decoherence that occurs when the Kolmogorov complexity of at least one of the involved systems approaches the Kolmogorov complexity of the external observer. In this case Y cannot maintain a complete description of X interacting with S and must discard some of the degrees of freedom. If we assume that all human observers acting in their SIA capacity have approximately the same Kolmogorov complexity, this situation will provide an explanation of the fact that we never see a human observer (or, say, a cat) as a quantum system.

4. Elements of reality

4.1 Entropic criterion of objectivity

Ever since the Einstein-Podolsky-Rosen article (1935), the question of what is real in the quantum world has been at the forefront of all conceptual discussions about quantum theory. The original formulation of this question involved physical *properties*: e.g., are position or momentum real? This is however not the only problem of reality that appears when many observers enter the game. Imagine a sequence of observers X_i, $i = 1, 2, \ldots$, each identifying systems S_n, $n = 1, 2, \ldots$. System identifications of each S_n do not have to coincide as some observers may have their Kolmogorov complexity $K(X_i)$ below, or close to, $K(S_n)$, and others much bigger than $K(S_n)$. If there is disagreement, is it possible to say that the systems are real, or objects of quantum mechanical investigation, in some sense? We can encode the binary identification string produced by each observer in his SIA capacity as some random variable $\xi_i \in \Omega$, where Ω is the space of such binary identification strings, possibly of infinite length. Index i is the number of the observer, and the values taken by random variable ξ_i bear index n corresponding to "i-th observer having identified system S_n". Adding more observers, and in the limit $i \to \infty$ infinitely many observers, provides us with additional identification strings. Putting them together gives a stochastic process $\{\xi_i\}$, which is an observation process by many observers. If systems S_n are to have a meaning as "elements of reality", it is reasonable to require that no uncertainty be added with the appearance of further observers, i.e., that this stochastic process have entropy rate equal to zero:

$$H(\{\xi_i\}) = 0. \tag{2}$$

We also take this process to be stationary and ergodic so as to justify the use of Shannon entropy.

Let us illustrate the significance of condition (2) on a simplified example. Suppose that $\theta_1, \theta_2, \ldots$ is a sequence of independent identically distributed random variables taking their values among binary strings of length r with probabilities q_k, $k \leq 2^r$. These θ_k can be seen as identifications, by different SIAs, of different physical systems, i.e., a special case of the ξ_i-type sequences having fixed length and identical distributions. For instance, we may imagine that a finite-length string, θ_1, is a binary encoding of the first observer seeing an electron and θ_2 is a binary string corresponding to the second observer having identified a physical system such as an elephant; and so forth. Then entropy is written simply as:

$$H = -\sum_k q_k \log q_k. \tag{3}$$

Condition (2) applied to entropy (3) means that all observers output one and the same identification string of length r, i.e., all SIAs are identical. This deterministic system identification, of course, obtains only under the assumption that the

string length is fixed for all observers and their random variables are identically distributed, both of which are not plausible in the case of actual quantum mechanical observers. So, rather than requiring identical strings, we impose condition (2) as a criterion of the system being identified in the same way by all observers, i.e., it becomes a candidate quantum mechanical "object of investigation".

4.2 Relativity of observation

Let us explore the consequences of condition (2). Define a binary sequence α_n^i as a concatenation of the system identifications strings of systems S_n by different observers:

$$\alpha_n^i = \overline{(\xi_1)_n}\,\overline{(\xi_2)_n}\ldots\overline{(\xi_i)_n}, \tag{4}$$

where index i numbers observers and the upper bar corresponds to "string concatenation" (for a detailed definition see Zvonkin and Levin 1970). Of course, this concatenation is only a logical operation and not a physical process. A theorem by Brudno (1978, 1983) conjectured by Zvonkin and Levin (1970) affirms that the Kolmogorov complexities of strings α_n^i converge towards entropy:

$$\lim_{n\to\infty}\lim_{i\to\infty}\frac{K(\alpha_n^i)}{i} = H(\{\xi_i\}). \tag{5}$$

For a fixed i and the observer X_i who observes systems S_n that are quantum in the sense of Definition 2, variation of $K(\alpha_n^i)$ in n is bounded by the observer's own complexity in his SIA capacity:

$$K(\alpha_n^i) < K(X_i) \qquad \forall n, \qquad i \text{ fixed}. \tag{6}$$

Hence eqs. (2) and (5) require that

$$\lim_{i\to\infty}\frac{K(\alpha_n^i)}{i} = 0. \tag{7}$$

This entails that the growth of $K(\alpha_n^i)$ in i must be slower than linear. Therefore the following:

Proposition 3. *An element of reality that may become an object of quantum mechanical investigation can be defined only with respect to a class of not very different observers.*

To give an intuitive illustration, imagine adding a new observer X_{i+1} to a group of observers $X_1, \ldots, X_i$ who identify systems S_n. This adds a new identification string that we glue at the end of concatenated string α_n^i consisting of all X_i's identifications of S_n, thus obtaining a new string α_n^{i+1}. The Kolmogorov complexity of α_n^{i+1} does not have to be the same as the Kolmogorov complexity of α_n^i; it can grow, but not too fast. Adding a new observation may effectively add some new non-compressible bits, but not too many such bits. If this is so,

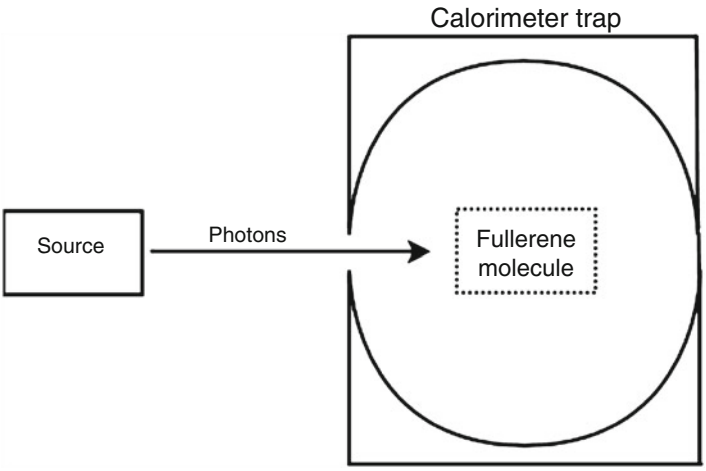


Figure 1: Experiment leading to heat production when observer's memory becomes saturated.

then $H = 0$ still obtains. Although observers $X_1, \ldots, X_i, X_{i+1}$ produce slightly different identification strings, they will agree, simply speaking, that an atom is an atom and not something that looks more like an elephant.

The above reasoning applies only to quantum systems S_n in the sense of Definition 2. This is because, in the case of non-quantum systems, different observers may operate their own coarse-graining, each keeping only some degrees of freedom. System identification strings may then differ dramatically, and one cannot expect $K(\alpha_n^i)$ to grow moderately.

5. Experimental test

A previously suggested experimental connection between thermodynamics and theories based on Kolmogorov complexity is based on observing the consequences of a change in the system's state (Zurek 1989, 1998; Erez et al. 2008). Zurek (1989) introduced the notion of physical entropy $\mathcal{S} = H + K$, where H is the thermodynamic entropy and K the Kolmogorov entropy. If the observer with a finite memory has to record the changing states of the quantum system, then there will be a change in $\mathcal{S}$ and it will lead to heat production that can be observed experimentally. We propose here a test independent of the change of state.

An individual fullerene molecule is placed in a highly sensitive calorimeter and bombarded with photons, which play the role of quantum systems with low $K(S)$ (Figure 1). The fullerene is a SIA, or a quantum mechanical observer, with $K(X) > K(S)$. Thus the absorption of the photon by the fullerene can be described as measurement: the fullerene identifies a quantum system, i.e. the photon, and observes it, obtaining new information. Physically, this process amounts

to establishing a correlation between the photon variables (its energy) and the vibrational degrees of freedom of the fullerene. From the point of view of an observer external to the whole setting, the disappearance of the photon implies that the act of observation by the fullerene has occurred, although the external observer of course remains unaware of its exact content.

Informationally speaking, the same process can be described at storing information in the fullerene's memory. If measurement is repeated on several photons, more such information is stored, so that at some point total Kolmogorov complexity of concatenated identification strings will approach $K(X)$. When it reaches $K(X)$, the fullerene will stop identifying incoming photons as quantum systems. Any further physical process will lead to heat production due to memory erasure, as prescribed by Landauer's principle (Landauer 1961). Physically, this process will correspond to a change of state of the carbon atoms that make up the fullerene molecule: the calorimeter will register a sudden increase in heat when C_{60} cannot store more information, thereby ending its observer function.

Actual experiments with fullerenes show that this scenario is realistic. A fullerene molecule "contains so many degrees of freedom that conversion of electronic excitation to vibrational excitation is extremely rapid". Thus, the fullerene is a good candidate for a quantum mechanical observer, for "the molecule can store large amounts of excitation for extended periods of time before degradation of the molecule (ionization or fragmentation) is observed" (Lykke and Wurz 1992). The experiments in which fullerenes are bombarded with photons demonstrate that "the energy of the electronic excitation as a result of absorption of a laser photon by a molecule is rapidly converted into the energy of molecular vibrations, which becomes distributed in a statistical manner between a large number of the degrees of freedom of the molecule... The fullerene may absorb up to 10 photons at $\lambda = 308$ nm wavelength before the dissociation of the molecule into smaller carbon compounds" (Eletskii and Smirnov 1995). We read these results as a suggestion that there should be one order of magnitude difference between $K(S)$ and $K(X)$ and that this allows the fullerene to act as a quantum mechanical observer for up to 10 photons at 308 nm wavelength. What needs to be tested experimentally in this setting is heat production: we conjecture that if the same process occurs inside a calorimeter, the latter will register a sudden increase in heat after the fullerene will have observed 10 photons (Figure 2). What we predict here isn't new physics, but an explanation of a physical process on a new level: that of information. We suggest that heat production deserves special attention as a signature of the fullerene's role as quantum mechanical observer.

As a side remark, imagine that the photon's polarization state in some basis were fully mixed:

$$\frac{1}{2}(|0\rangle + |1\rangle).$$

While only the energy of the photon matters during absorption, the external observer records von Neumann entropy $H = \log 2$ corresponding to this mixture (the initial state of the fullerene is assumed fully known). After absorption, it is manda-

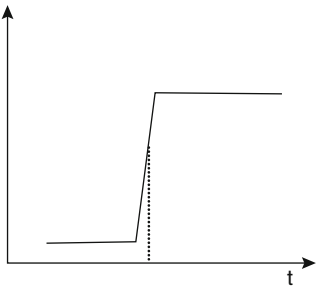


Figure 2: Conjectured time dependence of heat production in the calorimeter (vertical axis). A sharp increase occurs when the fullerene's memory is erased as it stops 'observing' photons quantum mechanically.

tory that this entropy be converted into Shannon entropy of the new fullerene state, corresponding nicely to the uncertainty of the external observer in describing the "statistical manner" of the distribution over a large number of the degrees of freedom. From the internal point of view, we may assume perfect 'self-knowledge' of the observer, which puts his Shannon entropy equal to zero. However, his Kolmogorov entropy will increase as a result of recording the measurement information (Zurek 1998). Heat produced during the erasure of measurement information is at least equal to the Kolmogorov complexity of the string that was stored in observer's memory; but, according to quantum mechanics, this heat will not reveal to the external observer any information about the precise photon state observed by the fullerene.

6. Concluding remarks

Information-theoretic treatment of quantum mechanical observer provides a formal result that encapsulates the Einstein-Podolsky-Rosen notion of "element of reality". We have shown how to make sense of a system existing independently of observation, with respect to a class of observers whose Kolmogorov complexities may differ, even if slightly. Equation (7) provides a mathematical criterion. It remains an open problem to find out whether the information-theoretic definition of observer will yield useful insights in other areas of quantum mechanics. We are currently pursuing this research program for studying quantum mechanical non-locality.

Acknowledgements: I am grateful to Vasily Ogryzko for stimulating discussions and to Časlav Brukner, Markus Aspelmeyer, Ognyan Oreshkov and Anton Zeilinger for their remarks and hospitality at the Institute for Quantum Optics and Quantum Information of the Austrian Academy of Sciences.

References

Bastin, T., (Ed.), 1971, *Quantum Theory and Beyond*. Cambridge: Cambridge University Press.

Bohm, D., 1971, "On Bohr's Views Concerning the Quantum Theory", in: T. Bastin (Ed.), 1971, p. 33.

Bohr, N., 1934, *Atomic Theory and the Description of Nature*. Cambridge: Cambridge University Press. Quoted in [**?**].

Brudno, A., 1978, "The Complexity of the Trajectories of a Dynamical System", in: *Russian Mathematical Surveys* 33, 1, pp. 197-198.

Brudno, A., 1983, "Entropy and the Complexity of the Trajectories of a Dynamical System", in: *Trans. Moscow Math. Soc.*, 2, pp. 157-161.

Einstein, A., Rosen, N., and Podolsky, B., 1935, "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?", in: *Physical Review* 47, pp. 777-780.

Eletskii, A. V., and Smirnov, B. M., 1995, "Fullerenes and Carbon Structures", in: *Physics-Uspekhi* 38, 9, pp. 935-964.

Erez, N., Gordon, G., Nest, M., and Kurizki, G., 2008, "Thermodynamic Control by Frequent Quantum Measurements", in: *Nature* 452, pp. 724-727.

Everett, H., 1957, "'Relative State' Formulation of Quantum Mechanics", in: *Review of Modern Physics* 29, pp. 454-462.

Friedman, M., 2001, *Dynamics of Reason*. Stanford: CSLI Publications.

George, A. (Ed.), 1953, *Louis de Broglie, physicien et penseur*. Paris: Albin Michel.

Jammer, M., 1974, *The Philosophy of Quantum Mechanics*. New York: John Wiley and Sons.

Kolmogorov, A., 1965, "Three Approaches to the Definition of the Concept 'Quantity of Information"', in: *Probl. Inform. Transm.* 1, 1, pp. 3-7.

Landau L., and Lifshitz, E., 1977, *Quantum Mechanics*. Pergamon Press.

Landauer, R., 1961, "Irreversibility and Heat Generation in the Computing Process", in: *IBM Journal of Research and Development* 5, pp. 183-191.

London, F., and Bauer, E., 1939, *La théorie de l'observation en mécanique quantique*. Paris: Hermann.

Lykke K. R., and Wurz, P., 1992, "Direct Detection of Neutral Products from Photodissociated C_{60}", in: *Journal of Physical Chemistry* 96, pp. 3191-3193.

Peres, A., 1986, "When Is a Quantum Measurement?", in: *American Journal of Physics* 54, 8, pp. 688-692.

Rovelli, C., 1996, "Relational Quantum Mechanics", in: *International Journal of Theoretical Physics* 35, p. 1637.

von Neumann, J., 1932, *Mathematische Gründlagen der Quantenmechanik*. Berlin: Springer.

Wigner, E., 1961, "Remarks on the Mind-Body Question", in: I. Good (Ed.), *The Scientist Speculates*, London: Heinemann, pp. 284-302.

Wigner, E., 1983, "Interpretation of Quantum Mechanics. Lectures Given in the Physics Department of Princeton University in 1976", in: J. A. Wheeler and W. Zurek (Eds.), *Quantum Theory and Measurement*. Princeton: Princeton University Press, pp. 260–314.

Zurek, W., 1989a, "Algorithmic Randomness and Physical Entropy", in: *Physical Review A* 40, pp. 4731-4751.

Zurek, W., 1989b, "Thermodynamic Cost of Computation, Algorithmic Complexity and the Information Metric", in: *Nature* 341, pp. 119-124.

Zurek, W., 1998, "Decoherence, Chaos, Quantum-Classical Correspondence, and the Algorithmic Arrow of Time", in: *Physica Scripta*, T76, 1998, pp. 186-198.

Zvonkin, A., and Levin, L., 1970, "The Complexity of Finite Objects and the Development of the Concepts of Information and Randomness by Means of the Theory of Algorithms", in: *Russian Mathematical Surveys* 25, 6, pp. 83-124.

CEA-Saclay, SPEC/LARSIM
91191, Gif-sur-Yvette Cedex
France
alexei.grinbaum@cea.fr

Leon Horsten

Mathematical Philosophy?[1]

Abstract

This article reflects on the scope and limits of mathematical methods in philosophy.

1. Introduction

Open a journal in chemistry at an arbitrary page, and you will see formulae. And it is the same for almost every other scientific subject. These formulae indicate that the article draws upon mathematics. It is like that if you open an issue of a journal that has just come out. But it is also like that if you open an issue of a journal that has appeared 50 years ago. Now open a general journal in philosophy at an arbitrary page. Chances are that you will see no formulae, but just text. This tells you that the article you are looking at does not draw on mathematical techniques or theories: it is written in a discursive style. This will certainly be the case if you open an an issue of a general philosophy journal that appeared 50 years ago. It is very likely also to be the case if you open a recent issue. Thus a discipline like chemistry is said to be a technical subject, whereas philosophy is said to be a non-technical subject.

This situation is currently changing rapidly. Until fairly recently, mathematical methods were used only in certain relatively specialised areas of philosophy such as philosophy of mathematics and philosophy of science. But in the last two decades, mathematical methods have become increasingly used in traditional areas of philosophy (such as epistemology and metaphysics).

It is time for a methodological reflection on this evolution. Until now, such a methodological investigation has not been carried out as far as I know. There has been some discussion on the use of mathematical models and methods in the philosophy of science ((van Benthem 1982), (Horsten and Douven 2008), (Muller t.a.), (Leitgeb t.a.), (Wheeler t.a.)). But there has been almost no systematic discussion of the use of mathematical methods in core areas of philosophy such as

1 This article is based on my inaugural lecture, which I gave in Bristol in December 2010. Thanks to Richard Pettigrew, Hannes Leitgeb, Neil Coleman, and Gregory Wheeler for valuable comments on earlier drafts of this article, and for stimulating conversations on the subject. Research for this article was partially supported by the AHRC project "Foundations of Structuralism" (AH/H001670/1).

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_7,

metaphysics and epistemology. Moreover, the methodological debate has focussed mainly on the use of *logical* methods.

This article is a systematic philosophical investigation into the role of mathematical methods in core areas of philosophy. I want to reflect on the scope and limits of mathematical methods in philosophy. I will argue that, while there are limits to what we can expect of mathematical methods in philosophy, mathematical methods can make a contribution to philosophy. I will not try to prove my case by giving you an annotated list of success stories and telling you what is so great about them. Instead, I want to look behind the examples. How is it that mathematical methods can contribute to philosophy? Which parts of philosophical inquiry necessarily have to be carried out in the discursive style?

I will not offer concrete methodological advice here: the concrete dangers and pitfalls of bringing mathematical methods to bear on philosophical problems have been discussed elsewhere ((Rota 1988), (Hansson 2000), (Horsten and Douven 2008)). And I also don't want to spend much time on the history of the use of mathematical methods in philosophy: a brief overview of this can be found in (Horsten and Pettigrew 2011).

2. Logical analysis and logical explication

Over the centuries it has occasionally been suggested that philosophy and mathematics are intimately related. Think of Plato's admonition "let no one destitute of geometry enter my gates", Spinoza's ideal of conducting philosophical research *more geometrico*, Leibniz' slogan "sedeamus et calculemus", etc. These sentiments were mainly fuelled by a craving for absolute certainty in philosophy. But in practice, philosophy has for most of its history been a discursive practice. This only slowly started to change in the beginning of the 20^{th} century. Ironically, this change happened at the time when the search for apodictic certainty started to lose its grip on philosophy.

In their logical investigations of mathematics, Frege and Russell recognised that the *logical* form of certain sentences differs in important ways from their surface grammatical form ((Frege 1879), (Russell 1905)). Russell and the early logical positivists emphasised that this was in particular the case for many philosophical propositions. By bringing out the logical form of philosophical propositions, certain supposedly deep philosophical problems could be unmasked as *pseudo-problems*. For a while the opinion held sway in some circles that in this manner, all philosophical propositions will, after logical analysis, turn out to be either trivially true, or trivially false, or empirical. In other words, it was thought that after logical analysis, there would be no deep philosophical questions left.

This turned out not to be the case. Most of the age-old central philosophical problems turned out to be impossible to dismiss as pseudo-problems, even after logical analysis. Only philosophical problems that were somewhat suspect in the first place – they sounded a bit silly – could be dissolved by logical analysis. So the ambitions of logical analysis had to be scaled back.

Carnap thought that logic nonetheless had an important role to play everywhere in philosophy. In his view, philosophers should strive at giving *logical explications* of philosophical concepts (Carnap 1950). When presented with a philosophical proposition, the philosopher should first of all uncover its logical form (as in the method of logical analysis). This logical form will relate certain predicates, and is given in a formal language. Next, Carnap says, plausible basic *rules of use* should be spelled out. The aim is to decide the philosophical proposition by deriving it or its negation from the rules of use using the rules of logic alone. Of course there is absolutely no guarantee that we will be able to solve the problem in this way: our basic principles may be controversial, or they may simply be too weak to decide the philosophical question we are interested in.

The method of logical explication has been criticised by the ordinary language philosophers on various grounds. One of their critiques rests on the fact that we cannot jump outside natural language. They argue that it is an illusion to think that appeal to formal languages can be a decisive step forward if one wants to address a philosophical question. Suppose that we are indeed able to derive the philosophical proposition we were interested in from correct rules of use. Then this whole derivation can be translated back to natural language, and one would not have had to make the detour via formal languages in the first place (Strawson 1963).

This much is true: there is no formal substitute for philosophical thinking. By working meticulously in formal languages invalid argumentative leaps can be excluded, and the possibility of tacit assumptions that remain under the radar is eliminated. But if one is careful enough, this can also be achieved in ordinary language. In any case, the philosophical argumentation will centre around the question of the basic rules of use that will be proposed. And this is a discussion that will take place in informal English. In sum, for all that has been said so far, while methodologically useful, semi-mathematical tools such as formal languages do not touch the heart of philosophy: they are dispensable in principle.

3. The dawn of mathematics in philosophy

In the late 1920s, Carnap started using mathematical models in philosophy. His ambitions were high: he wanted to construct the whole world (!) in terms of elementary sensory data and a similarity relation between those data. His models were set-theoretic in nature. In this programme, the colour red, for instance would turn out to be something like a set of sets of . . . sensory data. And even physical space would turn out to be some such set. There is no need to go into the details of Carnap's "logical construction of the world" (Carnap 1928), because it was ultimately unsuccessful. Instead, let us look at a use of models that is generally regarded as at least somewhat successful.

The first kind of non-set-theoretical models that were used in philosophy are probabilistic models. Probabilistic models were the first examples of quantitative

models in philosophy. Such models were used to shed light on the problem of confirmation in the philosophy of science. Suppose you have a scientific theory, and suppose you obtain a new piece of observational evidence. Then this piece of evidence can confirm or disconfirm the theory. Philosophers of science wanted to articulate a satisfactory philosophical theory of this support-relation between theories and evidence.

It turned out to be very hard for philosophers to find tenable basic principles of confirmation using Carnap's method of logical explication. The reason why the method of logical explication did not produce satisfactory results is that our intuitions about the confirmation relation are unreliable. By the beginning of the 1950s it had become clear that whenever we list principles concerning the confirmation relation that agree with our intuitions, they are invariably met with counterexamples. Somehow it seemed that a theoretical idea was needed.

In the late 1950s, philosophers of science took up the idea of modelling confirmation as probability-raising: evidence confirms a hypothesis if it raises the hypothesis' probability. This was the start of developing probabilistic models for studying the confirmation relation. Note that this development does not fit well with Carnap's method of logical explication: it is hard to imagine that the concept of probability somehow belongs to the logical form of all propositions involving the confirmation relation.

A whole machinery (known as Bayesian confirmation theory) has since then been developed to tackle problems in confirmation theory. And whatever one may think of it, this research programme was more successful than the approach to confirmation that came before, which was a version of Carnap's method of logical explication.

The probabilistic models contributed to our understanding of the confirmation relation by giving us an understanding of our intuitions concerning confirmation (Earman 1992). Before the advent of probabilistic models, we knew that our intuitions in this area are unreliable. But we did not really understand why. Probabilistic models provide compelling and integrated stories of why and in which situations our confirmation intuitions go astray. They show how our intuitions are shaped and sometimes deceived by our experience. The probabilistic models reintegrate and organise our intuitions.

For a model construction programme such as Carnap's (Carnap 1928), the ideal aim could be to find the unique correct model: the way the world actually is built up from experience. But the subjective probability-approach to confirmation never really aimed at uniqueness. From the start, the prior assignments of probability values were taken to be somewhat arbitrary, and were taken to irredeemably vary from person to person. Thus their theory was fundamentally a large ensemble of models rather than a unique intended model. This is very much in consonance with the model-theoretic or semantic view of theories in the natural sciences.

4. Recent uses of mathematical methods in philosophy

So probabilistic modelling plays an important role in the sub-discipline of philosophy of science that is called confirmation theory, whereas, at least until its very recent revival (Leitgeb 2007), the idea of set-theoretically constructing the world from experience was seen as a lost cause. There were a few other areas in philosophy where mathematical modelling played some role (such as philosophy of mathematics). But these are all somewhat specialised and relatively new areas of philosophy. Philosophy of science, for instance, came to its own in the first decades of the 20th century. In the core and more traditional areas of philosophy, such as general epistemology, metaphysics, and ethics, mathematical modelling was not done at all. And the mathematical methods used were in some sense 'logical'. Set theory is a part of mathematical logic, and some say that probability theory is somehow a 'generalised' form of logic.

This situation has begun to change in the past two decades. To an ever larger extent, mathematical modelling, as well as other mathematical techniques, are used even within traditional, core areas of philosophy. And the techniques and models that are used draw upon a large variety of mathematical fields (graph theory, mathematical analysis, algebra,. . .).

Let me mention some examples from epistemology and metaphysics. A fundamental epistemological question is: *why should our credences satisfy the standard laws of probability?* In recent work starting with (Joyce 1998), techniques and results of mathematical analysis have been used in the formulation and exploration of proposed answers to this question that involve distance-minimalisation. (De Clercq and Horsten 2005) have invoked techniques of graph theory to formulate identity conditions for secondary qualities such as colour shades. More examples could be listed, but instead I want to discuss one example (from metaphysics) in some more detail – of course this will be a highly simplified account.

Nominalists believe that the world, absolutely all there is, consists of concrete objects that stand in a part-whole relation to each other. Abstract objects *do not* exist, according to nominalism.

Nominalism is a philosophical theory if there ever was one. It is a metaphysical doctrine, dating back at least to the Middle Ages. But modern-day nominalists are enough of a naturalist to want their theory to be compatible with empirical science. The theories of the modern natural sciences use mathematics. So nominalism somehow has to find a way of recognising the truth of key principles of mathematics. Let us concentrate on the theory of the natural numbers: that is surely a key and basic mathematical theory.

There are two obstacles for the nominalist here. First, number theory seems at first blush about abstract entities. After all, in which museum is the number 7 held? Secondly, there is the question whether the world of the nominalist is large enough to accommodate the natural numbers. There are infinitely many numbers: who knows if there are infinitely many concrete objects?

In response to the first problem, it seems that the nominalist has to let concrete objects somehow play the role of the natural numbers: concrete objects is all she's got![2] In response to the second problem, the nominalist has to bite a bullet, and assume the existence of infinitely many concrete objects. This is perhaps not completely hopeless if space-time is infinite in some dimension – perhaps the time-dimension in the future direction.

From the 1950s onwards, nominalists (such as Nelson Goodman) have thought about precise principles governing the part-whole relation. In the 1960s, a minimal theory was gradually settled on, together with a list of possible extra principles that might also be true, but that are not universally accepted in the nominalist community (Niebergall 2011).

Now the following question emerges. Given that there are enough concrete objects to stand proxy for the natural numbers, can the basic axioms that govern the natural numbers somehow be validated? Roughly, this means the following: can the language of arithmetic somehow be translated into the nominalistic language of concrete objects and the part-whole relation, in such a way, that the basic principles of arithmetic are validated? In the light of the foregoing, it should be clear that from the present-day nominalist point of view, this is an elementary question that is of utmost importance. If it can be done, then a nominalist understanding and recognition of the laws of elementary arithmetic is possible.

Somewhat surprisingly, the answer turns out to be 'no' (Niebergall 2011). It has been shown in the past two decades that the nominalistic theories, minimal or extended by further principles that have been advocated, cannot validate even 'minimal' arithmetical theories. For the *cognoscendi*: they cannot even interpret the arithmetical system known as Robinson arithmetic, which is standard arithmetic without the axiom of mathematical induction. The proofs of this are in fact not really difficult: they only involve some relatively elementary facts about Boolean algebras.

Has the philosophical question of the viability of nominalism thereby been settled in the negative? Has a philosophical problem been laid to rest? No. For it is open to the nominalist to change her position and to say that not just the part-whole relation, but other, more complicated relations between concrete objects are nominalistically acceptable as belonging to the basic fabric of the world. When that is done, we have moved to the next round in the debate about nominalism.

But the point is that we will have advanced: we have made progress in this philosophical debate. We have not solved the question of nominalism; but we have shed light on it. And, coming back to the ordinary language philosophers, it is hard to see how this insight could have been obtained using the discursive methodology of ordinary language philosophy. In principle, Niebergall's proofs about the noninterpretability of Robinson Arithmetic in standard mereological theories can

2 An alternative for the nominalist is to develop a fictionalist position concerning mathematical objects. (Thanks to Neil Coleman for pointing that out.) But here I assume that indispensability arguments justify adopting a realist line on the question of the existence of mathematical objects.

be spelled out in ordinary English, just like any mathematical proof can. But it is unreasonable to think that Niebergall's arguments could have been produced in practice using the methods of ordinary language philosophy.

5. Limitations?

In the debate between ordinary language philosophy and 'formal' philosophy, objections against the methodology of logical explication have crystallised only gradually. The application of a wide variety of mathematical methods to central problems in philosophy is a very recent phenomenon. The opposition hasn't had time yet to get organised. In the years to come, that will probably happen. What follows is a glimpse of what their arguments might look like.

5.1 Philosophy and our conceptual world

One objection of the ordinary language philosophers that will undoubtedly re-emerge, is naturalistic in spirit. On the one hand there are the objects, properties, and relations in the world. It is the business of the sciences to describe what is out there in the world; philosophy had better not try to compete with them. On the other hand, there are our everyday concepts and conceptions, which latch imperfectly onto the world. It is the business of philosophy to describe our concepts: this is called *conceptual geography*. Our concepts are a fairly loose and gerrymandered lot. Now you can, using the process of logical explication, find substitutes for these concepts that are more structured, in the sense that they satisfy a small set of highly coherent and general basic principles. But when you have arrived at these substitutes, you have lost contact with our concepts as we live them in our experience. In Rota's words:

> The concepts of philosophy are among the least precise. The mind, perception, memory, cognition, are words that do not have any fixed or clear meaning. Yet they do have meaning. We misunderstand these concepts when we force them to be precise. (Rota 1988, p. 170)

This is not to say that there is anything wrong with trying to find mathematical structure in our conceptual world.[3] But it is somewhat unlikely that our conceptual world is mathematically structured in the way in which the physical world miraculously has turned out to be. And if it isn't, then it's no use pretending that it is. This is what Wittgenstein had in mind when he said that as soon as philosophy has produced a theory, you can bet on it that it is wrong (Wittgenstein 1956).

This is a deep and important point. The relation between ordinary language philosophy and phenomenology, on the one hand, and our concepts and our experience on the other hand, is somewhat like the relation between literary criticism

3 My former colleague Hannes Leitgeb emphasises that this is a valid objective of mathematical philosophy.

and literature. Literary criticism is somehow continuous with literature; discursive philosophy is continuous with our conceptual world. All of it belongs to our culture and will do so in the future. Moreover, discursive philosophy is not just an epiphenomenon of the culture and society we live in. It changes our conceptual world and our lived experience.

Mathematical philosophy has more in common with the particular part of our culture that we call science, which is much less continuous with our everyday conceptual world. Mathematical philosophy wants to play with the hard-hitting girls. Its ambition is not just to describe our concepts, but to capture properties and relations in the world. That is a tough proposition, but it seems to me that there is no way around it.

This is also where Carnap's insistence that the concepts that play a role in the logical explication must be *fruitful* is relevant. Carnap emphasised that the rule of use-principles need not be a completely faithful representation of the way in which the concepts involved are used in ordinary language. But, Carnap says, these formal substitutes for our ordinary concepts somehow have to be theoretically useful. And it is in this context that we should understand Strawson's critique of Carnap, when he says that Carnap's method is like offering someone a book on physiology when she asks (with a sigh): "who understands the human heart?" (Strawson 1963).

The hard-headed position that I am advocating here does not exclude that some of the properties that the mathematical philosopher wants to investigate, are subject-relative in some way. To take an example, consider confirmation again. As we have seen, many philosophers of science now think that the confirmation relation contains a subjective component. Nonetheless, the philosophers of science aim at more than describing our concept or concepts of confirmation; they aim at describing what it means for evidence to confirm a hypothesis.

It may be that in some areas, attempts to go beyond the geography of our everyday concepts are doomed to fail. Suppose, for instance, that not only all attempts to 'uncover the grounds of morality' turn out to be futile, but that even all attempts to derive most accepted moral maxims from a small and coherent number of principles fail. (This may, in so far as I can see, actually be the case.) This would then be an area where mathematical methods could never be applied fruitfully in the way that Carnap envisaged.

5.2 Models and instrumentalism

The following is often seen as an obstacle to playing with the big girls. In the natural sciences, sensory experience (observation and experimentation) is our ultimate touchstone. Theories are tested on the basis of their empirical consequences. Philosophical theories are also connected with sensory experience, but in a much less definite way, and their connection with the outcome of scientific experiments

is even less clear.[4] How do we refute a philosophical theory, even if it is precisely formulated (as a class of models, say)?

It is often said that our common sense intuitions form the touchstone of philosophical theories.[5] The value of this depends on what the philosopher's aims are. If she wants to be faithful to our concepts and the relations between them, then our intuitions indeed occupy a privileged position. But if her aim is to latch on to an 'objective' relation or property, then our intuitions may well be unreliable – although they are unlikely to be even then massively mistaken: there is often deeper truth hidden behind intuitive falsehoods.

Indeed, success of a philosophical way of picturing the phenomena is not easy to define. It is a matter of shedding light on a subject, of providing insight, of showing how it all hangs together. The paucity of precise empirical predictions does not bar philosophy from obtaining objective knowledge. As Alonzo Church once said: the preference of *seeing* over *understanding* as a method of observation seems capricious (Church 1951). In other words, there may well be situations in which philosophers have good reasons to believe in the objective correctness of models that they produce: the key factor will be explanatory power.

This takes us to a fundamental difference between the role of models in the natural sciences and in philosophy. In the natural sciences, models can be valuable even if they are fundamentally unrealistic, not in the sense of making idealisations (such as the absence of friction), but in the sense of *intentionally* making fundamentally false incorrect assumptions (as is done for instance in modelling traffic as a fluid passing through a system of connected tubes). Even though such models do not really *explain* anything, they serve an important goal: they are connected to observational and experimental predictions. Even models that do not describe the world anywhere near correctly can be extremely powerful as a source of empirical predictions. Indeed, even an empiricist such as van Fraassen who is agnostic about the existence of unobservable entities, properties, and relations, is happy to acknowledge the value of models that postulate sub-atomic particles. An instrumentalist stance to models is always possible in science.

Philosophical theories do not typically make precise empirical predictions. Thus if one does not believe in the objective correctness of a class of models in philosophy (even granting the idealisations involved), then its value is much less clear. As intimated earlier, the way in which one can bring oneself to believing in the objective correctness of a class of models is in philosophy basically the same as in the sciences. It consists in success arguments. Ultimately, they are variants of *Inference to the Best Explanation*. Nonetheless, even classes of models in philosophy that are perhaps difficult to take seriously, such as the part-whole models discussed earlier, may have their value. They function as a conceptual

4 This point is emphasized in (Hansson 2000).

5 There is also the question who is meant with 'our' in this sentence. Experimental philosophers hold that many of the 'intuitions' on which analytical philosophy is built are generated by a quite unrepresentative sample of the population, and therefore suspect. I will leave this discussion aside here.

laboratory (van Benthem 1982). They give us insight in what metaphysically might have been, in a way in which theories of magnetic monopoles give us insight in what physically might have been.

5.3 Informal concepts and the discursive style

Classes of mathematical models are built using very precise concepts. This causes classes of mathematical models to have a certain *rigidity*: it is difficult to adapt classes of mathematically defined models to phenomena that they were not intended to describe in the first place. Classes of models that are generated by a mathematical technique are also very *stubborn*. Once a certain mathematical modelling technique has firmly taken hold of a field, it is very difficult to replace it with a new mathematical modelling technique or just to get rid of it if it doesn't work well. Genuinely new classes of mathematical models that are suitable for describing phenomena in a given field are very difficult to find.

Even though he was a great advocate of the use of models, Boltzmann pointed out that everyday concepts possess a *plasticity* that scientific concepts to a large extent lack (Boltzmann 1902).[6] Everyday concepts are in a sense like stem cells: they can become virtually anything. This is definitely a virtue when we are working in an area where we are still groping for clues, when we are still feeling our way around. In such a situation, the discursive method is the only way, for we still have to shape our concepts. And, as we know from medical science, it is important that we always keep a healthy supply of stem cells. You never know when you will need them. At some point, we may have to look for a new class of models, and then we simply have to start with our everyday concepts.

In some situations, stem cells take on a sharply defined shape: they commit themselves to a specific task and agree to a division of labour. This corresponds to the emergence of sharply defined models in science and in philosophy. At its best, models can have the effect of 'switching on the light'; at their worst they merely serve as the intellectual equivalent of wearing blinders. In any case, they are a prerequisite for having a theory that can really be tested. Precisely because models are precise, and somehow rigid, and somehow narrow-minded, they cannot easily dodge attempts at refutation.

Nonetheless, here again a difference with the use of models in the natural sciences emerges. Let us return for a moment to the example of nominalism. We have seen that the decision to take only the part-whole relation as fundamental can be and has been challenged. The debate about the correctness of taking the part-whole relation as the only basic relation is conducted in the discursive style. And this is in large part where the philosophical action is. More in general, philosophical disputes about the form and basic ingredients of the models must be conducted in the discursive style. There is no other way: conducting the discussion in the language of the model would beg the question. In this way, the discursive style

6 Frege also made this point (Frege 1879, introduction).

necessarily forms a constitutive part of any philosophical investigation. In the natural sciences, discussion about the basic ingredients of the models are less central. Again, this has to do with the fact that observational evidence forms the ultimate touchstone. Many scientists believe that as long as a class of models yields the right empirical predictions, there can be no legitimate cause for concern or criticism. Even if one does not believe that, there is much less at stake. As I have said before, even unrealistic models can be of utmost importance in science.

5.4 The bounded scope of mathematical methods

In the light of all this, we may ask: what then are the virtues of using logico-mathematical methods? Where is the pay-off?

Carnap's method of logical explication forces one to make the grammar and the structure of philosophical argumentation explicit. This is obviously a good thing: it is a question of intellectual hygiene. But its instances do not import an essential use of mathematical methods in philosophy. Rota puts it too harshly, perhaps, when he writes:

> Confusing mathematics with the axiomatic method for its presentation is as preposterous as confusing the music of Johann Sebastian Bach with the techniques for counterpoint in the Baroque age. (Rota 1988, p. 171)

In the case of nominalism, we have seen how mathematical methods can really enter into it. They can be used to prove limitative results, or *impossibility results* as they are sometimes called. In the case of part-whole nominalism, the principles turn out too be too weak to do significant mathematics.

But there are also situations where the principles we come up with are too strong. Think about the case of the theory of truth, where the liar paradox teaches us that intuitive basic truth principles lead to a contradiction. As a response to this, philosophers have tried to weaken the truth principles in such a way that the basic intuition behind them is still preserved as much as possible. In order to show that these weakened principles are at least consistent, one has to produce a model in which they are true. In this way, models can yield *possibility results*. Of course when one has produced a model, one has only a mathematical possibility result, and this falls far short of showing that the theory under investigation is a serious philosophical contender. Again, to substantiate the latter claim, a discursive story has to be told.

Mathematical models, such as probabilistic models in the case of confirmation, can *unify* a seemingly disparate array of intuitions. Carnap's method of logical explication can do this to some extent, but use of mathematical models and techniques are much more powerful in this respect. One reason for this is that a class of models can show what binds a collection of basic principles together, more so than a list of axioms can. A mathematical class of models gives us a way of looking at a class of phenomena in a unified way.

Models are ways of looking at something. Sometimes one can look at a phenomenon in different ways that are in some sense equally fruitful. Take the case of subjunctive conditional sentences: sentences of the form

If A were the case, then B would be the case.

One can look at subjunctive conditionals in a probabilistic way. That is, one can say (roughly) that a conditional sentence of that type are true (or acceptable) if and only if $Pr(B \mid A)$ is high. But one can also look at them in a 'topological' way. That is, one can say (roughly) that a conditional sentence of that type is true if and only if the situations in which A is true that are 'close' to the way things actually are, are also situations in which B is true. Now there are *representation theorems* which show (roughly) that for every 'probabilistic' model for subjunctive conditionals, there exists a 'topological' model that is equivalent to it, and, conversely, that for every 'topological model', there is a 'probabilistic' model that is equivalent to it (Leitgeb unpubl.). With equivalence is meant here that they classify the same subjunctive sentences as true. In other words, mathematical theorems can sometimes tell us that there is a sense in which two different ways of looking at something nevertheless in some sense yield the same results.

So the picture I want to suggest is the following. At the beginning, we have a philosophical hypothesis, informally expressed. In this form, its content is to a degree fluid and indeterminate. In order to understand the hypothesis, and eventually to assess it, we have to make it more definite and more precise. This can be achieved by associating with it a class of mathematical models. (This can of course be done in more than one way.) Only then mathematical techniques and results come into play. They allow us to *understand* the content of the models. They increase our insight into an interpretation of the philosophical hypothesis with which we started.

Nonetheless, there are limits to the power of mathematical methods in philosophy. As an essential but proper part of a philosophical account, mathematical models and methods can help shed light on philosophical problems. But even supposing that deep philosophical problems can in principle be solved: what mathematical methods can never ever do, is to single-handedly solve philosophical problems. This can never happen. For the reasons that I have given, philosophical theories will always remain more closely connected to our informal concepts and to our informal way of arguing than theories from the natural sciences. It would be folly to think that the discursive style of informal philosophy can ever be eliminated in any branch of philosophy. Use of mathematical methods will never be a substitute for philosophical thought.

References

Boltzmann, L., 1902, *Model.* Entry in the Encyclopedia Brittanica.

Carnap, R., 1928, *Der logische Aufbau der Welt.* Felix Meiner Verlag.

Carnap, R., 1950, *The Logical Foundations of Probability.* University of Chicago Press.

Church, A., 1951, "The Need for Abstract Entities in Semantical Analysis", in: *American Academy of Arts and Sciences Proceedings* 81, pp. 110-133.

De Clercq, R., and Horsten, L., 2005, "Closer", in: *Synthese* 146, pp. 371-393.

Earman, J., 1992, *Bayes or Bust?* Cambridge (Mass.): The MIT Press.

Frege, G., 1879, *Begriffsschrift. Eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.* Louis Nebert.

Hansson, S., 2000, "Formalization in Philosophy", in: *Bulletin of Symbolic Logic* 2, pp. 162-175.

Horsten, L., and Douven, I., 2008, "Formal Methods in the Philosophy of Science", in: *Studia Logica* 89, pp. 151-162.

Horsten, L. and Pettigrew, R., 2011, "Mathematical Methods in Philosophy", in: L. Horsten and R. Pettigrew (Eds.), *Continuum Companion to Philosophical Logic.* Continuum Press, pp. 14-26.

Joyce, J., 1998, "A Nonpragmatic Vindication of Probabilism", in: *Philosophy of Science* 65, pp. 575-603.

Leitgeb, H., 2007, "A New Analysis of Quasi-analysis", in: *Journal of Philosophical Logic* 36, pp. 181-226.

Leitgeb, H., "Logic in General Philosophy of Science: Old Things and New Things", in: *Synthese*, to appear.

Leitgeb, H., *A Probabilistic Semantics for Counterfactuals.* Unpublished manuscript, 2010.

Müller, T., 2010, "Formal Methods in the Philosophy of Natural Science", in: F. Stadler (Ed.), *The Present Situation in the Philosophy of Science.* Springer.

Niebergall, K.-G., 2011, "Mereology", in: L. Horsten and R. Pettigrew (Eds.), *Continuum Companion to Philosophical Logic.* Continuum Press.

Rota, J.-C., 1988, "The Pernicious Influence of Mathematics upon Philosophy", in: *Synthese* 88, pp. 165-178.

Russell, B., 1905, "On Denoting", in: *Mind* 14, pp. 398-401.

Strawson, P., 1963, "Carnap's Views on Constructed Systems versus Natural Languages in Analytical Philosophy", in: P. A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap*, pp. 503-518.

van Benthem, J., 1982, "The Logical Study of Science", in: *Synthese* 51, pp. 431-452.

Wheeler, G., 2012, "Formal Epistemology", in: A. Cullison (Ed.), *Continuum Companion to Epistemology.* Continuum Press.

Wittgenstein, L., 1956, *Philosophical Investigations.*

Department of Philosophy
University of Bristol
43 Woodland Road
BS8 1UU, Bristol
UK
Leon.Horsten@bristol.ac.uk

Ulrike Pompe

The Value of Computer Science for Brain Research

Abstract

The intrinsic relationship between computer science and brain research fuels a number of philosophically interesting questions. The present essay focuses on two major aspects of this relationship: the enabling role of computer science for brain research on the one hand and the use of computational means to simulate or re-build the brain on the other hand. Even though these two streams of thought are distinct their combination helps to elucidate a deeper problem (or so I hope), namely the question what it is exactly that we can find out by rebuilding the brain.

1. Introduction

The intrinsic relationship between computer science and brain research is not only long standing and widely acknowledged, it is as well interesting from a philosophical point of view: on the one side there are philosophers of science who find an historically interesting example of the importance of suitable analogies for the progress of science and on the other side philosophers of mind will find this analogy itself interesting since the way research is conducted from the perspective of the brain-computer comparison provides new insights into the scopes and limits of the analogy, helping us to refine our understanding of the concept of mind. The present essay focuses on two major aspects of the brain-computer relationship: the enabling role of computer science for brain research on the one hand and the use of computational means to simulate or re-build the brain on the other hand. In order to be able to appreciate the role of computer science for the examination of the brain, namely in its role of providing a new and utile analogy for how the brain works, I will present a (very) brief walk through the history of brain research, thereby supporting the thesis that the computer and the principles of information processing served as an essential metaphorical analogy to the brain's functioning. Concerning the second aspect, I will introduce a recently started large-scale research project, the so-called Blue Brain Project, and comment on its advertised goals. Even though these are two quite distinct lines of thought, combining them helps us to elucidate a deeper problem (or so I hope), namely the question what it is exactly that we are able to find out about the brain by rebuilding or simulating it. I believe that some expectations concerning the results of reverse-engineering approaches, so-called whole-brain simulations, as announced by the founders of

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_8,

the Blue Brain Project, will probably not be met – the support for my skepticism is not derived from general skepticism concerning the technical realization but from an already older debate in the philosophy of mind, namely the debate about cognitivism vs. the so-called 4-E movement which proposes that the restricted focus on the brain without consideration of the periphery, such as the body, the environment, societal and behavioral motivations and constraints, will fail to explain the specificities of the human mind. I close by advocating for another, complementary use of simulations within brain research, sketching thus an alternative to (or further development of) the described whole-brain simulations. This part, I'm afraid, is more a tentative idea than a full-fledged proposal.

2. Brain research and its need for analogies

Most (his)stories[1] about the beginning of modern brain research start with Descartes. Descartes, inspired by the mechanical inventions of his contemporaries, compared the human body with a mechanical machine,[2] all parts of which are subject to mechanical laws. The nerve fibers are compared to tiny tubes through which the so-called *spiritus animalis*, small particles which transmit the driving force for bodily action, are flowing. They originate from the seat of the soul, the pineal gland and push from there through the liquid in the nerve tubes to those parts of the body that are to be activated, e.g. to the arm when it is supposed to be raised. The pushing of the liquid in the muscles and the nerve tubes is thus a hydraulic-mechanical force. In this sense, the machine "body" was thus thought to be governed by the immaterial soul and the interface between body and soul was placed in the pineal gland, because it is the only part in the brain which does not exist in lateralized, i.e. twofold, form.

Under this premise, namely that the brain is the ultimate seat of the soul, the brain moved to the center of interest in science and philosophy. What followed was an increase in anatomical and physiological studies of the human and animal brain. During this early period of brain research, throughout the 17th and 18th century, a great number of anatomical and physiological studies were more or less systematically conducted. Among other things, it was discovered that nerve fibers are irritable but not sensitive themselves; that tendons are neither sensitive nor irritable, therefore that they are not nerves; that there is grey and white matter; that there are several small chambers within the brain filled with liquid; that the brain stem regulates breathing and heart rate and that the brain itself is insensitive.[3] However, the project of finding the seat of the soul by empirical means came

1 See for example: a) Erhard Oeser, *Geschichte der Hirnforschung*. Darmstadt: Wissenschaftliche Buchgesellschaft 2002. b) Michael Hagner, *Homo Cerebralis – Der Wandel vom Seelenorgan zum Gehirn*. Frankfurt/Main und Leipzig: Insel Verlag 2000.

2 Compare René Descartes, de Homine (1622); Traité de l'homme (1664).

3 Compare Oeser, *Geschichte der Hirnforschung, op.cit.,* pp. 58-101.

slowly to an end as, despite the vast efforts in anatomical and physiological study, it remained a deep puzzle how the interaction between a non-material entity like the soul and material entities like the brain and the body came about. Silently, this research endeavor went out of steam.

The next note-worthy development in brain research, though taking a different approach, was Gall's program called "*Organologie*".[4] Gall was interested in human character traits and how they might be visible by means of outward features. He believed to have found certain correlations between the form of the skull and the strength or weakness of certain character traits in individuals. His student, Spurzheim, took this program onto the next level, claiming that every character trait, like love for one's children, sense for beauty, truthfulness, humor, sense of justice, etc. occupied a certain part of the brain, the brain itself being just a collection of these different kinds of "organs" (hence: *Organology* or *Phrenology*) and that careful measuring of the skull could provide insight into the underlying brain structure and therefore, the character of the person. As doubtful as this program appears today, it provides a first localization project of the brain and its functions and thereby cleared the path for a very promising research agenda, namely the localization of distinct cognitive abilities as functions of circumscribed cortical areas. Parallel to Gall, and already way before, physicians had observed that head injuries could cause the impairment of cognitive abilities like speaking and discovered a variety of other more general disorders of mind. The first systematic observations were made by Paul Broca and Carl Wernicke[5] in the 19th century, who both studied patients suffering from speech loss or speech disorder. The posthumous autopsies revealed brain lesions in distinct neocortical areas; further research confirmed that the production (Broca-Area) and the auditory decoding (Wernicke-Area) of speech are correlated with the well-functioning of these areas. Once the idea that certain areas of the neocortex, the outer and most voluminous part of the brain, perform or realize mental and cognitive ability had found acceptance, the puzzle of how the brain actually did this had arisen and began to dominate research. The research had thereby moved from describing the brain in physiological and anatomical terms to finding out about its functional setup. What are the ultimate constituents of the brain and how do they perform? At the beginning of the 20th century, answering these questions seemed so far out of reach – possibly due to the lack of a suitable analogy with the brain in nature and technology – that psychologists restricted their work to what was observable, namely behavior. The aim of behaviorism[6] was thus to study the stable patterns of stimulation by environmental cues and the resulting behavior. A human mind – just like that of most other animals – was considered to be an input-output device, the inner workings of which were not important. This "black box" of the mind was the big challenge

4 Compare Oeser, *Geschichte der Hirnforschung, op.cit.,* pp. 110-130.

5 Compare Oeser, *Geschichte der Hirnforschung, op.cit.,* pp. 157-165; also: Brian Kolb and Ian Q. Wishaw, *Neuropsychologie*. Heidelberg: Spektrum Verlag 1996.

6 Burrhus Frederic Skinner, *About Behaviorism*. Vintage 1974.

for brain research – a challenge science was not able to tackle until the rise of a suitable analogy. It is here where the rise of computer science helped the progress of psychology and added a new dimension to brain research.

3. Computer Science as the way out of the black box

When Alan Turing and his contemporaries proposed the first digital computation machines, the "toolkit" for understanding neuronal activity was created. The mathematical foundation of an information processing machine which could be programmed in multiple ways provided a simple and elegant model with which the brain, at that time still conceived an input-output device, could be compared. A further enabling discovery for the new brain sciences was the major discovery of the synapse, by Donald Hebb in 1949,[7] which accounted for the functioning of interneural connection. In a nutshell, he found out that "what fires together, wires together", indicating that neural activity strengthens the connection between the involved neurons, whereas the lack of activity between neurons disassembles them. With these discoveries (or developments), the basic principles of information processing and learning had been established, and not only on the behavioral level, but on a smaller, significantly organic scale. Indeed, it did not take long until these principles were exploited for the creation of the first artificial neurons. Already in 1943, McCulloch and Pitts[8] had developed the first computational neuron: it consisted in a kind of logical element which had a threshold mechanism and several entries leading to one single exit. This exit can enter the states TRUE or FALSE. The TRUE state is achieved when the sum of the entry signals exceeds the threshold; if not, the element remains in the FALSE state. These properties are analogous to those of an action potential of a neural cell. Further properties of neural cells have been modeled this way. In 1952, for example, the so-called Hodgkin-Huxley Model[9] of a neuron has been presented. It describes the electrical properties of the cell membrane, thereby allowing to model action potentials of nerve cells. Finally, only a couple of years between Turing's seminal paper in 1936[10] and Hebb's findings, the first artificial neural net was created, the Percep-

7 Richard E. Brown and Peter M. Milner, "The legacy of Donald O. Hebb: More than the Hebb Synapse", in: *Nature Reviews Neuroscience*, 4, 2003, pp. 1013-1019.

8 Warren S. McCulloch and Walter H. Pitts, "A logical calculus of the ideas immanent in nervous activity", in: *Bulletin of Mathematical Biophysics*, 5, 1943, pp. 115-133.

9 Alan L. Hodgkin and Andrew F. Huxley, "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve", in: *Journal of Physiology*, 117, 1952, pp. 500-544.

10 Alan Turing, "On Computable Numbers, With an Application to the Entscheidungsproblem", in: *Proceedings of the London Mathematical Society*, Series 2, 42, 1936; reprinted in M. David (Ed.), *The Undecidable*, Hewlett, NY: Raven Press 1965.

tron by Rosenblatt in 1958.[11] It combined a small number of "digital" neurons and ordered them in a set of fixed layers. The Perceptron was thereby able to "compute" simple functions like AND, NOT, and OR, thus covering a range of dissociative, associative and inhibiting functions.

These developments in mathematics and computer sciences provided a breakthrough for brain science: finally a useful analogy was found to capture what neural activity and thereby brain function comes down to. The firing of actual neurons, which is so hard to observe in the living system, could now be replaced by on and off states of simple neuronal nods, the combination of which produced a set of states which was able to represent a variety of informational states of the overall system. What followed in the decades after was a veritable explosion of scientific approaches to the brain and its functions. Parallel to the classical psychological research paradigms such as behavioral and psychophysical studies, psychopathological and post-mortem studies, a couple of invasive recording techniques evolved, such as single cell and cluster recordings of neural activity, global recordings like EEG and finally imaging techniques like PET and fMRI. Diverse methods and research paradigms have led to a patchwork-like research field in which a unified account of the brain and its overall function often is still missing. A vital part in the creation of testable models in which at least a few of the heterogeneous data sources can be recombined can be achieved by the use of artificial neural networks. The field of Computational Neuroscience, which began to evolve since the early 1980ies, sets its goal on providing a platform for "the theoretical study of the brain used to uncover the principles and mechanism that guide the development, organization, information processing and mental abilities of the nervous system",[12] thereby representing a methodological bridge between the different kinds of physiological, anatomical, and behavioral knowledge. Out of the same spirit, namely in order to unify the existing yet diverse research results from physiology, anatomy and other fields, the Blue Brain Project was created.

4. Simulating the brain: The Blue Brain Project

The above sketched problem concerning the patchwork-like knowledge of the brain that we gain from the multiplicity of research approaches and methods might be overcome if these results could be integrated into one single artificial brain. This is the core idea of the so-called reverse engineering approach which characterizes the Blue Brain Project.[13]

11 Frank Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", in: *Psychological Review*, 65, 1958, pp. 386-408.

12 Thomas P. Trappenberg, *Fundamentals of Computational Neuroscience*. Oxford: Oxford University Press 2005, p. 7.

13 Henry Markram, "The Blue Brain Project", in*: Nature Reviews Neuroscience*, 7, 2006, pp. 153-160.

The Blue Brain Project was founded in 2005 by Henry Markram at the École Polytechnique Fédérale de Lausanne (Brain Mind Institute) with the idea of building a virtual brain with the ultimate aim to understand how the brain functions:

> … the aims of this ambitious initiative are to simulate the brains of mammals with a high level of biological accuracy and, ultimately, to study the steps involved in the emergence of biological intelligence.[14]

The project has two distinct aims, it seems: it shall provide a powerful tool for a computational simulation of the biological organ and furthermore, it also targets to explain the organ's biological function: the realization of (intelligent) behavior. I believe, the first goal is an interesting and uncontroversial one and I will only briefly sketch how Markram and colleagues are tackling it. In a second step I want to come back to the second aim or target because it is here that I see a conceptual problem.

Concerning the realization of the project, Markram builds on already existing models of neurons and neuronal circuits, such as the Hodgkin-Huxley model of a neuron, together with Wilfred Rall's detailed model of dendritic and axonal arborizations of neurons[15]. These fundamental properties of individual neural cells can be combined to form artificial microcircuits and to model basic information processing mechanisms between neural cells, like lateral inhibition and feedback. The next step is to model different cell types (e.g. pyramidal cells) and build greater and greater compartments with an increasing degree of complexity. [16]

The first step for the Blue Brain Project is to recreate a so-called neocortical column. These columns are the basic building blocks of the neocortex. In the rat's neocortex, which serves as the model and data source for the project, these columns are usually composed of approximately 10,000 neocortical neurons. A column measures about 0.5 mm in diameter and about 1.5 mm in height. The data concerning the number, kind and connectivity of the neurons in a typical neocortical column is derived from recordings of neural cells in the somatosensory cortex of 14 to 16 day old rats. These columns are at that time not overly specialized so that they can still be generalized to other parts of the cortex, allowing a certain kind of flexibility within the simulation:

> the young sensory column, which is evolutionarily one of the simplest and is also highly accessible, is an ideal starting point from which the column can be 'matured' to study development, 'transformed' to study regional specialization and 'evolved' to study evolution of the neocortex.[17]

---

14 Markram, "The Blue Brain Project", *op.cit*., p. 153.

15 Wilfrid Rall, "Branching Dendritic Trees and Motoneuron Membrane Resistivity", in: *Experimental Neurology*, 1, 1959, pp. 491-527.

16 Compare Markram, "The Blue Brain Project", *op.cit*., Ref. 14-32.

17 Markram, "The Blue Brain Project", *op.cit*., p. 156.

This way, a template column is created – the so-called Blue Column which, just like a biological entity, will include different types of neurons in the distinct layers that the cortex consists of (six in sum).

As nicely as the technical details of implementation and data generation are laid out, the following statements give rise to some philosophical concerns. Markram writes: "Blue Brain will allow us to challenge the foundations of our understanding of intelligence and generate new theories of consciousness".[18] In a BBC World Service interview, Markram even goes so far as to commit to the statement that "if we build it correctly it should speak and have an intelligence and behave very much as a human does."[19] These statements bring us back to the second and to my notion more problematic envisioned goal of the project.

The sort of intelligence Markram is interested in is described as a qualitatively new element which emerges from the dynamical interaction of *electrical "molecules"*. Markram compares the emergence of a new qualitative level of mental properties to a number of similar phenomena in which complex structures emerge from the regular combination of simpler base entities. The observed "qualitative leaps"[20] arise from the increasing complexity and can only be understood if the rules according to which they emerge out of the properties of the base entities are understood. Molecules, for example, can be described by investigating their special anatomical layout. A chain of increasingly complex "phenomena" or entities could thus look like this: atoms – molecules – DNA molecules – genes – proteines – cells – cell types – brain regions. Each scale for itself is not sufficient to express the complexity of the higher scale, which induces Markram to claim that the next higher scale possesses a different kind of quality. This reasoning leads him to believe that once we understand how the essential building blocks of the brain interact we will find out about intelligence because this phenomenon seems to be the next scale:

> The brain seems to make the next quantum leap in the quality of intelligence, beyond the physical structures to form dynamic electrical 'molecules'. The ultimate question, therefore, is whether the interaction between neurons drives a series of qualitative leaps in the manner in which information is embodied to represent an organism and its world.[21]

The proposed multi-scaling approach to biological complexity as proposed here is not critical; it is a reasonable approach to describing multiple levels of organic life. The question is whether the phenomenon of *intelligence* can be easily included into the hierarchy. For one, the concept is underspecified here – if it is meant to exclusively capture information processing, then we would have to subsume a

18 Markram, "The Blue Brain Project", *op.cit*., p. 159.

19 Jonathan Fildes (22 July 2009). "Artificial brain '10 years away'". BBC News. http://news.bbc.co.uk/2/hi/8164060.stm.

20 Compare Markram, "The Blue Brain Project", *op.cit*., p. 153.

21 Markram, "The Blue Brain Project", *op.cit*., p. 153.

great number of computational operations under it, which do not meet our intuitive requirements for ascribing intelligence. If it is meant to capture information processing and also information filtering for the realization of flexible, adaptive and purposeful behavior, then, however, the notion of information processing does not suffice to capture intelligence. We need at least include aspects like motivation, decision, deliberation, and other factors that mark intelligent behavior. This is concern number one: behavior ranges on a much larger scale in temporal and spatial dimensions than information processing. Simulating information processing, even if it is captured in all its complexity, might not elucidate what intelligence is, unless this concept is properly defined. What this boils down to is that the concept of intelligence might not range in the class of phenomena that can be reduced to information processing alone.[22] This point leads to the second concern, namely whether a pure focus on brain activity will ever suffice to account for the mental, but this will be discussed further below. For the moment it is worth having a look at the proposed reverse-engineering approach and its expected merits: the immanent thesis is that consciousness and intelligent behavior (thus intelligence?) is a function of complexity. If we reverse-engineer the component parts and learn to understand their interaction then we learn something about behavior and consciousness and intelligence. We thus consider the building blocks and let them interact as if they were placed in a natural environment. However, when Markram claims that the elucidation of human intelligence will be made possible by the reverse engineering approach, he does not tell us how he will make the system "behave". The ensemble of neocortical columns would have to be embedded into a larger architecture: first, obviously in a kind of sensory environment, then in a kind of body, which will lead the system to have a "self", a bounded extension, which allows it to distinguish between itself and the environment. Without this, how could it act on and react towards external stimuli? A further point: a system does not act if it doesn't have a purpose, goal, aim, whatever you want to call it. These motivational factors are provided by elementary needs, such as for food, warmth, energy, etc. Where will these be derived from? How can we expect anything from a purely passive system? There will have to be an interface to make the system do something. How is the sensory environment going to be connected to and implemented into the artificial brain?

These might be technical issues, seen from the surface, but I believe there is something deeper to it. The data we use to model this environment cannot be

22 For an interesting discussion on the (im)possibility to define concepts, see Edouard Machery, "Why I stopped worrying about the definition of life … and why you should as well", in: *Synthese*, 185, 2012, pp. 145-164. My view is compatible with Machery's, since the argument does not hinge on a specific definition of the concept of intelligence, or a definition which is valid in both scientific and ordinary speech, rather, the argument pleads for a categorical difference between kinds of behavior (e.g. intelligent behavior vs. merely reflexive reactions) and scales of matter, such as atoms, molecules and organized units of these basic entities such as cells and organs.

derived from measurement – what is required is a theoretical model of the world we live in and we need an understanding or at least a hypothesis about how a human being is part of and is interacting with its physical and social environment. Simulations as such and the possibility to simulate whole brains provide a new tool that reopens an old theoretical debate. The issue centers around the question of whether we can explain "the mental" solely on the basis of modeling neural and core cognitive processes like the underlying neural mechanisms involved in perception, motor behavior, speech, memory, and ultimately – and this is the tricky part – thought. Some opponents to this restrictive cognitivist approach to the human mind claim that we will never be able to understand the specificity of mental properties, such as their contents, never be able to understand behavior if we do not consider the relevant motivational factors which trigger it; never be able to understand the constituents of cognitive acts if we don't consider the environmental vehicles on which cognitive acts rely and onto which they are externalized. The proponents of these claims[23] advocate for a view of the mind which emphasizes its embeddedness, its extendedness, embodiedness and also its enactive nature. The view they wish to challenge – in a nutshell – is that cognition is not all inside the skull. This so-called 4E movement (Embedded, Extended, Embodied, Enactive) is not essentially opposed to the classical cognitivist account – in fact, as Adams and Aizawa[24] point out, these positions are easily reconcilable. The general problem about the 4E contribution is that it is difficult to pinpoint how external constraints are exactly constituting certain mental events; the cognitivist approach also never denied that the brain is an embedded organ and that it persons who act and behave, not just the brain. However, to project this issue onto the Markram project, I believe there is a point here in raising these concerns.

What is it that we want to know when we are investigating the "big picture"? Certainly, the coming of existence of such a thing as consciousness is an interesting question and if a reverse-engineered brain can help to find out the basic foundation of consciousness, then it would be a welcome contribution. However – and this objection has been brought forth in recent debates in the philosophy of mind as stated above[25] – it is questionable whether "consciousness" (both its quality and its mere existence) can be explained by examining (or simulating) the organic structure of the brain. The brain does not do things unless it is stimulated. Stimulation can only occur if the organ is embedded in a periphery consisting of

23 See among others: a) Susan Hurley, *Consciousness in Action*. Cambridge MA: Harvard University Press 1998. b) Shaun Gallagher, *How the Body Shapes the Mind.* Oxford: Oxford University Press 2006. c) Alva Noë, *Action in Perception.* Cambridge (Mass.): The MIT Press 2006. d) Andy Clark and David Chalmers, "The Extended Mind", in: *Analysis*, 58, 1, 1998, pp. 7-19.

24 Kevin Adams and Ken Aizawa, "Embodied Cognition and the Extended Mind", in: P. Garzon and J. Symons, (Eds.), *Routledge Companion to the Philosophy of Psychology*. New York: Routledge 2009, pp. 193-213.

25 Compare footnote 22.

sense organs and a body and perhaps even a social and worldly environment. The explanative power of a purely bottom-up driven simulation approach might not reach beyond the boundaries of already known properties of cells and their interaction on the physiological level. It would take a complementary approach which includes behavioral data to allow a real explanation of why and when brain states and neural processes result in certain behavioral patterns.

5. Bottom-up vs. top-down simulations: Function before structure

The alternative I would like to sketch here is to complement the bottom-up simulation approach by an agent-driven approach. Behavioral data can be used to model a tentative underlying cognitive architecture (as it is practiced in cognitive science anyway), and if such a model could be implemented in a simulation of a virtual agent, where certain external and internal constraint factors can be varied independently of each other, then this simulation device would allow us to predict an agent's behavior under certain conditions, composed of external as well as internal constraints. The agent's behavior could be read out through reaction times and error rates, just as in classical behavioral experiments, the difference being total control over the parameters and a completely transparent agent. The internal constraints which might be implemented in the system could be e.g. global manipulations of neural activity like those prompted by sleep-deprivation or depression, or local manipulations like those resulting from small lesions. External variants might be stimulus strength and number, a set of tasks, etc. A detailed and biologically accurate model of the processes and constraints on the neural level – just as the Blue Brain Project provides – would then complete the picture. One would be able to explain observable behavioral patterns by the underlying neural constraints in dependence of external stimulation and internal information processing and information validation.

6. Conclusion

In the introduction to a special issue on Computation and Cognitive Science, Mark Sprevak states that

> the computational framework has rendered theorizing about inner processes respectable, it has provided a unified and naturalistic arena in which to conduct debates about psychological models, and it provides the tantalizing possibility of accurately simulating and reproducing psychological processes. There is almost universal agreement that the mind is in some sense like a computer. But consensus quickly ends once we ask *how*.[26]

26 Mark Sprevak, "Introduction to the Special Issue Computation and Cognitive Sci-

With the short review of the history of brain research I intended to show that there are multiple facets of research aims which lead to investigating the brain and that these goals firstly changed over time, but that they also center around one crucial question, namely what it is that accounts for the mental. Understanding the brain alone on the basis of its anatomical and physiological structures cannot account for any of its functions; neither does the pure observation of behavioral patterns provide a satisfactory account of the mental. The possibility to model simple neural nets definitely helped to shape a new, helpful analogy, namely that of the brain as a computational (thus information processing) device, thereby helping to find a new level of explanation of mental events and their constituents: the exploitation of this analogy provided a deep and sound understanding of neural connectivity and its virtues for cognition. However, philosophers raised the objection that the focus on the brain and on neural activity alone will not suffice for solving the more global puzzle. This global puzzle of the mental demands not only a thorough description of the organic structures but also a functional account as to why and when certain processes occur while also taking into account peripheral constraints such as the social and bodily environment.

It may be senseless to ask after the ultimate goal of brain research since we are facing a multi-scale phenomenon. However, the explanation and clarification of underspecified concepts as that of consciousness and others will in the end require the united efforts of firstly, theoretical disciplines which reflect on the scope and meaning of this term and secondly, empirical disciplines, which investigate the physiology of the brain on the one hand, but, importantly, its functional embedding in a human being under the perspective of its behavioral, social, and cognitive demands. If large and multi-scale simulations can be employed to do so, then they provide a valuable contribution. It is doubtful, however, that these means – in a purely bottom-up sense – will qualitatively supersede the more classic experimental approaches and be able to replace them in the long run.

Philosophy of Simulation
Institute of Philosophy
University of Stuttgart
Seidenstrasse 36
70174, Stuttgart
Germany
Ulrike.Pompe@philo.uni-stuttgart.de

ence", in: *Studies in the History and Philosophy of Science* , 41, 2010, pp. 223-226, here cited p. 223.

SAM SANDERS

ON ALGORITHM AND ROBUSTNESS IN A NON-STANDARD SENSE

ABSTRACT

In this paper, we investigate the invariance properties, i.e. robustness, of phenomena related to the notions of algorithm, finite procedure and explicit construction. First of all, we provide two examples of objects for which small changes completely change their (non)computational behavior. We then isolate robust phenomena in two disciplines related to computability.

1. INTRODUCTION

The object -or better concept- of study in Computer Science is unsurprisingly *computation*. The notions of *algorithm*, *finite procedure* and *explicit computation* are central. The present paper investigates the robustness of these notions, i.e. we are interested in phenomena regarding computation which are reasonably stable under variations of parameters. Let us first consider two illuminating examples of *non-robust* phenomena in Computer Science.

Example 1. Recently[1] the following remarkable mathematical object was developed: a pair of *computable*[2] random variables (X, Y) for which the conditional distribution $P[Y|X]$ is *non-computable*[??], as it codes the *Halting Problem*[??]. Let CAM be the statement that such (X, Y) exists. Before trotting out all sorts of indispensability claims based on CAM, one should bear in mind that the conditional distribution $P[Y|X]$ becomes *computable* again[3], after the addition to Y of some kind of generic noise E. Let CAM_E be the statement that $P[Y + E, X]$ is computable, for computable (X, Y) and generic noise E. Evidently, we may see CAM_E as a variation of CAM involving an error parameter. However, the (non)computational content of CAM is completely different from that of CAM_E. Indeed, the addition of the noise E dramatically changes the non-computability of $P[Y, X]$, and hence the computational content of CAM_E, compared to CAM. In short, the computational behavior of $P[Y|X]$ is sensitive to minor perturbations and CAM is non-robust with regard to the addition of error parameters.

1 See (Freer et al. 2011).

2 The words in italics have precise technical definitions to be found in e.g. (Soare 1987).

3 This explains why, in any real-world scenario invariably involving noise, the non-computability of $P[Y|X]$ never manifests itself.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_9,

Example 2. In Constructive Analysis[4], the notion of *finite procedure* is central. An object only exists after it has been *constructed* (in finitely many steps). The following is a well-known negative result of the constructive school: Given a uniformly continuous function on $[0, 1]$ such that $f(1) < 0$ and $f(0) > 0$, we cannot in general construct $x_0 \in [0, 1]$ such that $f(x_0) = 0$. In other words, the intermediate value theorem, INT for short, cannot be proved in Constructive Analysis. By contrast, we have the following positive result, called INT_E: Given $\epsilon \in \mathbb{R}$ and given a uniformly continuous function on $[0, 1]$ such that $f(1) < 0$ and $f(0) > 0$, we can construct x_0 such that $|f(x_0)| < \epsilon$. Again, we may see INT_E as a variation of INT involving an error parameter. Analogous to the previous example, the computational behavior of the intermediate value is sensitive to minor perturbations: the addition of an error term makes the former computable (in the sense of Constructive Analysis). In other words, INT also exhibits computational non-robustness with regard to the addition of error parameters.

The previous examples provide phenomena regarding computation that are destroyed by a minor variation. In this paper, we intend to identify phenomena regarding computation that are not affected by certain variations (like perturbation of parameters). In other words, we are looking for *robust* behavior in topics related to Computer Science. The importance of robustness cannot be overestimated, as our scientific models of reality are only approximations and tend to incorporate imprecise assumptions, often for valid reasons such as workability, elegance or simplicity. Thus, if a phenomenon X occurs in a robust model, we are reasonably certain that X cannot be ascribed to an artifact of the model, but corresponds to a real-world phenomenon X'.

A similar point has been made in the past by Ian Hacking and Wesley Salmon. In particular, the numerous independent ways of deriving Avogadro's constant (with negligible errors) are taken by Hacking and Salmon to be sufficient evidence for the real-world existence of molecules and atoms.[5] Another example from Hacking is concerned with the photo-electric effect.

> The simple inference argument says it would be an absolute miracle if for example the photoelectric effect went on working while there were no photons. The explanation of the persistence of this phenomenon [...] is that photons do exist. As J. J. C. Smart expresses the idea: 'One would have to suppose that there were innumerable lucky accidents about the behavior mentioned in the observational vocabulary, so that they behaved miraculously as if they were brought about by the non-existent things ostensibly talked about in the theoretical vocabulary.' The realist then infers that photons are real [...] (Hacking 1983, pp. 54-55).

In general, numerous independent derivations of the same phenomenon make it implausible that the latter is an artifact of a particular framework or modeling

4 See (Bishop 1967) and (Bridges and Vîţă 2006).

5 See (Hacking1983, pp. 54-55), (Salmon 1984, pp. 214-220) and (Salmon 1998, pp. 87-88).

assumption, i.e. the phenomenon in question is about something real[6]. Hence, by seeking out the robust phenomena involving computation, we may get a better understanding of the real core of computation, while at the same time develop a better theory of what exactly constitutes robustness.

We begin our search in two disciplines related to computability, *Reverse Mathematics* and *Constructive Analysis*, both introduced below. First, in Section 2, we study the invariance properties present in Reverse Mathematics, a discipline intimately connected to computability. Secondly, we do the same for Errett Bishop's constructive notion of algorithm from Constructive Analysis in Section 3.

2. Reverse Mathematics

In this section, we identify certain invariance properties in *Reverse Mathematics*. The latter is closely related to *Recursion Theory*, a classical framework for studying (non)computability. A central object in Recursion Theory is the *Turing machine*[7], introduced next.

2.1. Alan Turing's machine and Recursion Theory

In 1928, the famous mathematician David Hilbert posed the *Entscheidungsproblem*. In modern language, the Entscheidungsproblem (or 'decision problem') asks for no less than the construction of an algorithm that decides the truth or falsity of a mathematical statement. In other words, such an algorithm takes as input a mathematical statement (in a suitable formal language) and outputs 'true' or 'false' after a finite period of time.

Before the Entscheidungsproblem could be solved, a formal definition of algorithm was necessary. Both Alonzo Church and Alan Turing provided such a formalism,[7] being the λ-calculus and the Turing machine, respectively. Church showed that, if the notion algorithm is formalized using the λ-calculus, then the construction required to solve the Entscheidungsproblem is impossible. Independently, Turing showed that the Entscheidungsproblem can be reduced to the *Halting Problem*, which is known to have no algorithmic solution, assuming 'algorithm' is identified with 'computation on a Turing Machine'. In time, it was shown that both formalisms, though quite different in nature, enable the computation of the same class of functions, now called the *recursive functions*. The latter class was intended to formalize the notion of recursion, later giving rise to Recursion Theory.

6 In light of Examples 1 and 2, we may rest assured that intermediate values and conditional probabilities will always be computable *in practice*, as actual computational practice suggests.

7 See (Church 1936) and (Turing 1937). Intuitively, a Turing machine is an idealized computer with no limits on storage and meomory.

Because of the correspondence between these three formalisms, it is generally accepted that we should identify the (informal and vague) class of algorithmically computable functions with the class of function computable by a Turing machine. This identification hypothesis is called the *Church-Turing thesis*. However, as suggested by Example 1, not all computability phenomena are robust. In the next section, we identify a phenomenon in Reverse Mathematics which *is*.

2.2. Reverse Mathematics and robustness

Reverse Mathematics is a program in the Foundations of Mathematics founded[8] in the Seventies by Harvey Friedman. Stephen Simpson's famous monograph *Subsystems of Second-order Arithmetic* is the standard reference.[9] The goal of Reverse Mathematics is to determine the *minimal* axiom system necessary to prove a particular theorem of ordinary mathematics. Classifying theorems according to logical strength reveals the following striking phenomenon.[9]

> It turns out that, in many particular cases, if a mathematical theorem is proved from appropriately weak set existence axioms, then the axioms will be logically equivalent to the theorem.

This phenomenon is dubbed the 'Main theme' of Reverse Mathematics. A good instance of the latter may be found in the Reverse Mathematics of WKL_0[10]. An example of the Main Theme is that the logical principle WKL is equivalent to Peano's existence theorems for ordinary differential equations $y' = f(x, y)$, the equivalence being provable in RCA_0. Some explanation might be in order: the system RCA_0 may be viewed as the logical formalization of the notion Turing machine, which in turn formalizes the notion of algorithm.[11] The principle WKL (or *Weak König's Lemma*) states the existence of certain *non*-computable objects.[12]

We now consider the system[13] ERNA which has no a priori connection to RCA_0, or Reverse Mathematics, or computability[14]. We will show that a version of the Main Theme of Reverse Mathematics is also valid in ERNA, but with the predicate '=' replaced by '≈', i.e. equality up to infinitesimals from Nonstandard

8 See (Friedman 1975; 1976).

9 See (Simpson 2009) for an introduction to Reverse Mathematics and p. xiv for the quote.

10 See (Simpson 2009, Theorem I.10.3).

11 Thus, Reverse Mathematics is intimately tied to Recursion Theory and computability.

12 In particular, Weak König's Lemma states the existence of an infinite path through an infinite binary tree. Even for *computable* infinite binary trees, the infinite path need not be computable. In other words, WKL is false for the recursive/computable sets. See (Simpson 2009).

13 See (Sanders 2011) for an introduction to ERNA and a proof of Theorem 3.

14 In particular, ERNA was introduced around 1995 by Sommer and Suppes to formalize mathematics in physics. See (Sommer and Suppes 1996; 1997).

Analysis.[15] Indeed, the following theorem contains several statements, translated from (Simpson 2009, IV) into ERNA's language, while preserving equivalence.

Theorem 3 (Reverse Mathematics for ERNA + Π_1-TRANS)**.** *The theory* ERNA *proves the equivalence between* Π_1-TRANS *and each of the following theorems concerning near-standard functions:*

1. *Every S-continuous function on* $[0,1]$ *is bounded.*
2. *Every S-continuous function on* $[0,1]$ *is continuous there.*
3. *Every S-continuous function on* $[0,1]$ *is Riemann integrable*[16].
4. *Weierstrass' theorem: every S-continuous function on* $[0,1]$ *has, or attains a supremum, up to infinitesimals.*
5. *The strong Brouwer fixed point theorem: every S-continuous function* $\phi : [0,1] \to [0,1]$ *has a fixed point up to infinitesimals of arbitrary depth.*
6. *The first fundamental theorem of calculus:* $\left(\int_0^x f(t)\,dt\right)' \approx f(x)$.
7. *The Peano existence theorem for differential equations* $y' \approx f(x,y)$.
8. *The Cauchy completeness, up to infinitesimals, of* ERNA*'s field.*
9. *Every S-continuous function on* $[0,1]$ *has a modulus of uniform continuity.*
10. *The Weierstrass approximation theorem.*

A common feature of the items in the previous theorem is that strict equality has been replaced with $\approx$, i.e. equality up to infinitesimals. This seems the price to be paid for 'pushing down' into ERNA the theorems equivalent to Weak König's lemma. For instance, item (7) from Theorem 3 guarantees the existence of a function $\phi(x)$ such that $\phi'(x) \approx f(x,\phi(x))$, i.e. a solution, up to infinitesimals, of the differential equation $y' = f(x,y)$. However, in general, there is no function $\psi(x)$ such that $\psi'(x) = f(x,\psi(x))$ in ERNA + Π_1-TRANS. In this way, we say that the Reverse Mathematics of ERNA + Π_1-TRANS is a *copy up to infinitesimals* of the Reverse Mathematics of WKL_0, suggesting the following general principle[17].

Principle 4. *Let* $T(=)$ *be a theorem of ordinary mathematics, involving equality. If* RCA_0 *proves* $T(=) \Leftrightarrow \mathrm{WKL}$*, then* ERNA *proves* $T(\approx) \Leftrightarrow \Pi_1$-TRANS.

Furthermore, there are more results of this nature. In a forthcoming paper, we show examples of the following general principle[18].

15 For an introduction to Nonstandard Analysis, we refer to (Kanovei and Reeken 2004).

16 In ERNA, the Riemann integral is only defined up to infinitesimals.

17 A similar (and equally valid) principle is *If* $\mathrm{RCA}_0 \vdash T(=)$*, then* ERNA $\vdash T(\approx)$.

18 A similar principle is *If* $\mathrm{RCA}_0 \vdash T(=)$*, then* ERNA + Π_2-TRANS $\vdash T(\approxeq)$.

Principle 5. *Let* $T(=)$ *be a theorem of ordinary mathematics, involving equality. If* RCA_0 *proves* $T(=) \Leftrightarrow$ WKL*, then* ERNA + Π_2-TRANS *proves* $T(\approxeq) \Leftrightarrow$ Π_3-TRANS.

Here, the predicate '$\approxeq$' is best described as 'equality up to *arbitrarily small* infinitesimals'. At least two more variations[19] are possible and in each instance, we obtain a similar principle concerning equivalences.

We conclude that the equivalences proved in Reverse Mathematics display a certain degree of robust behavior: First of all, we observe similar series of equivalences in different frameworks. In other words, the equivalences observed in classical Reverse Mathematics are not an artifact of the framework, as they occur elsewhere in similar forms. Secondly, the equivalences in classical Reverse Mathematics remain valid when we consider different error predicates, i.e. replace equality by '$\approx$' or '$\approxeq$'. Thus, small perturbations in the form of error predicates do not destroy the observed equivalences.

3. Reuniting the antipodes

In this section[20], we show that the notion of algorithm in *Constructive Analysis* is endowed with a degree of robustness. This is achieved *indirectly* by defining a new notion called 'Ω-invariance' inside Nonstandard Analysis, and showing that it is close to the constructive notion of algorithm, as it gives rise to the same kind of Reverse Mathematics results. In other words, there are two different notions of finite procedure, i.e. the constructive notion of algorithm and Ω-invariance, which both give rise to the same kind of equivalences in (Constructive) Reverse Mathematics. Again, we observe that the latter are not affected by some change of framework.

3.1. The notion of finite procedure in Nonstandard Analysis

Here, we define Ω-invariance, a central notion, inside (classical) Nonstandard Analysis. We show that Ω-invariance is quite close to the notion of finite procedure.

With regard to notation, we take $\mathbb{N} = \{0, 1, 2, \dots\}$ to denote the set of natural numbers, which is extended to ${}^*\mathbb{N} = \{0, 1, 2, \dots, \omega, \omega + 1, \dots\}$, the set of *hypernatural* numbers, with $\omega \notin \mathbb{N}$. The set $\Omega = {}^*\mathbb{N} \setminus \mathbb{N}$ consists of the *infinite* numbers, whereas the natural numbers are *finite*. Finally, a formula is *bounded* or 'Δ_0', if all the quantifiers are bounded by terms and no infinite numbers occur.

19 The first one is the removing of parameters in Π_1-TRANS and the second one is the assumption of a greatest relevant infinite element.

20 The title of this section is explained in Remark 16 below. The italicized concepts are introduced in Section 3.2.

Definition 6 (Ω-invariance)**.** *Let $\psi(n, m)$ be Δ_0 and fix $\omega \in \Omega$. The formula $\psi(n, \omega)$ is* Ω-invariant *if*

$$(\forall n \in \mathbb{N})(\forall \omega' \in \Omega)(\psi(n, \omega) \leftrightarrow \psi(n, \omega')). \tag{1}$$

For $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, the function $f(n, \omega)$ is called Ω-invariant, if

$$(\forall n \in \mathbb{N})(\forall \omega' \in \Omega)(f(n, \omega) = f(n, \omega')). \tag{2}$$

Now, any object $\varphi(\omega)$ defined using an infinite number ω is potentially non-computable, as infinite numbers can code (non-recursive) sets of natural numbers.[21] Hence, it is not clear how an Ω-invariant object might be computable or constructive in any sense. However, although an Ω-invariant object clearly involves an infinite number, the object does not depend on the *choice* of the infinite number, by definition. Furthermore, by the following theorem, the truth value of $\psi(n, \omega)$ and the value of $f(n, \omega)$ is *already* determined at some finite number.

Theorem 7 (Modulus lemma)**.** *For every Ω-invariant formula $\psi(n, \omega)$,*

$$(\forall n \in \mathbb{N})(\exists m_0 \in \mathbb{N})(\forall m, m' \in {}^*\mathbb{N})[m, m' \geq m_0 \to \psi(n, m) \leftrightarrow \psi(n, m')].$$

For every Ω-invariant function $f(n, \omega)$, we have

$$(\forall n \in \mathbb{N})(\exists m_0 \in \mathbb{N})(\forall m, m' \in {}^*\mathbb{N})[m, m' \geq m_0 \to f(n, m) = f(n, m')].$$

In each case, the number m_0 is computed by an Ω-invariant function.

Proof. Although the proof of this lemma is outside of the scope of this paper, it is worth mentioning that it makes essential use of the fact that an Ω-invariant object does not depend on the *choice* of infinite number. □

The previous theorem is called 'modulus lemma' as it bears a resemblance to the modulus lemma from Recursion Theory.[22] Intuitively, our modulus lemma states that the properties of an Ω-invariant object are already determined at some *finite* number. This observation suggests that the notion of Ω-invariance models the notion of *finite procedure* quite well.

Another way of interpreting Ω-invariance is as follows: central to any version of constructivism is that there are basic objects (e.g. the natural numbers) and there are certain basic operations on these objects (e.g. recursive functions or constructive algorithms). All other objects are non-basic (aka 'non-constructive' or 'ideal'), and are to be avoided, as they fall outside the constructive world. It goes without saying that infinite numbers in ${}^*\mathbb{N}$ are ideals objects *par excellence*. Nonetheless, our modulus lemma suggests that if an object does not depend on the *choice* of ideal element in its definition, it is not ideal, but actually basic. This is the idea behind Ω-invariance: ideal objects can be basic if their definition does not really

21 See (Kreisler 2006).

22 See (Soare 1987, Lemma 3.2.)

depend on the choice of any particular ideal element. In this way, Ω-invariance approaches the notion of finite procedure *from above*, while the usual methods work *from the ground up* by defining a set of basic constructive operations and a method for combining/iterating these.

We now consider two examples of Ω-invariant objects.

Remark 8. First of all, assume we have $(\exists n \in \mathbb{N})\varphi(n)$, with $\varphi \in \Delta_0$. Then the function[23] $(\mu n \leq \omega)\varphi(n)$ is Ω-invariant. Hence, there is an Ω-invariant function providing a witness n_0 for $\varphi(n_0)$ (Compare item (5) in Definition 10).

Secondly, we show that a Δ_1-formula is Ω-invariant. To this end, assume $\psi \in \Delta_1$, i.e. for some $\varphi_1, \varphi_2 \in \Delta_0$, we have that

$$\psi(m) \leftrightarrow (\exists n_1 \in \mathbb{N})\varphi_1(n_1, m) \leftrightarrow (\forall n_2 \in \mathbb{N})\varphi_2(n_2, m), \tag{3}$$

for all $m \in \mathbb{N}$. Now fix some $\omega' \in \Omega$. Let $p_\psi(m)$ be the least $n_1 \leq \omega'$ such that $\varphi_1(n_1, m)$, if such exists and ω' otherwise. Let $q_\psi(m)$ be the least $n_2 \leq \omega'$ such that $\neg\varphi_2(n_2, m)$ if such exists and ω' otherwise. For $m \in \mathbb{N}$, if $\psi(m)$ holds, then $p_\psi(m)$ is finite and $q_\psi(m)$ is infinite. In particular, we have $p_\psi(m) < q_\psi(m)$. Now suppose there is some $m_0 \in \mathbb{N}$ such that $p_\psi(m_0) < q_\psi(m_0)$ and $\neg\psi(m_0)$. By (3), we have $(\forall n_1 \in \mathbb{N})\neg\varphi_1(n_1, m_0)$ and, by definition, the number $p_\psi(m)$ must be infinite. Similarly, the number $q_\psi(m_0)$ must be finite. However, this implies $p_\psi(m_0) \geq q_\psi(m_0)$, which yields a contradiction. Thus, we have $\psi(m) \leftrightarrow p_\psi(m) < q_\psi(m)$, for all $m \in \mathbb{N}$. It is clear that we obtain the same result for a different choice of $\omega' \in \Omega$, implying that ψ is Ω-invariant.

Care should be taken to choose the right axiom system to formalize the above informal derivation. Indeed, in certain axiom systems, not all Δ_1-formulas are Ω-invariant.

Finally, we consider the *transfer* principle from Nonstandard Analysis.

Principle 9. *For all φ in Δ_0, we have*

$$(\forall n \in \mathbb{N})\varphi(n) \to (\forall n \in {}^*\mathbb{N})\varphi(n). \tag{4}$$

The previous principle is called 'Π_1-TRANS' or 'Π_1-transfer'. Note that Π_1-transfer expresses that $\mathbb{N}$ and ${}^*\mathbb{N}$ have the same properties. In other words, the properties of $\mathbb{N}$ are *transferred* to ${}^*\mathbb{N}$. In what follows, we do not assume that this principle is given.

3.2. Constructive Analysis and Constructive Reverse Mathematics

In this section, we sketch an overview of the discipline *Constructive Reverse Mathematics* (CRM). In order to describe CRM, we first need to briefly consider Errett Bishop's *Constructive Analysis*.

23 The function $(\mu k \leq m)\psi(k)$ computes the least $k \leq m$ such that $\psi(k)$, for ψ in Δ_0. It is available in most logical systems.

Inspired by L. E. J. Brouwer's famous foundational program of intuitionism,[24] Bishop initiated the redevelopment of classical mathematics with an emphasis on *algorithmic* and *computational* results. In his famous monograph[24] *Foundations of Constructive Analysis*, he lays the groundwork for this enterprise. In honour of Bishop, the informal system of Constructive Analysis is now called 'BISH'. In time, it became clear to the practitioners of Constructive Analysis that intuitionistic logic provides a suitable formalization[25] for BISH.

Definition 10 (Connectives in BISH)**.**

1. The disjunction $P \vee Q$: we have an algorithm that outputs either P or Q, together with a proof of the chosen disjunct.
2. The conjunction $P \wedge Q$: we have a proof of P and of Q.
3. The implication $P \to Q$: by means of an algorithm we can convert any proof of P into a proof of Q.
4. The negation $\neg P$: assuming P, we can derive a contradiction (such as $0 = 1$); equivalently, we can prove $P \to (0 = 1)$.
5. The formula $(\exists x)P(x)$: we have (i) an algorithm that computes a certain object x, and (ii) an algorithm that, using the information supplied by the application of algorithm (i), demonstrates that $P(x)$ holds.
6. The formula $(\forall x \in A)P(x)$: we have an algorithm that, applied to an object x and a proof that $x \in A$, demonstrates that $P(x)$ holds.

Having sketched Bishop's Constructive Analysis, we now introduce Constructive Reverse Mathematics. In effect, Constructive Reverse Mathematics (CRM) is a spin-off from the Reverse Mathematics program introduced in Section 3.2. In CRM, the base theory is (inspired by) BISH and the aim is to find the minimal axioms that prove a certain *non-constructive* theorem. As in Friedman-Simpson style Reverse Mathematics, we also observe many equivalences between theorems and the associated minimal axioms.

We now provide two important CRM results.[26] First of all, we consider the *limited principle of omniscience* (LPO).

Theorem 11. *In* BISH*, the following are equivalent.*

1. LPO: $P \vee \neg P \quad (P \in \Sigma_1)$.
2. LPR: $(\forall x \in \mathbb{R})(x > 0 \vee \neg(x > 0))$.
3. MCT: (*The monotone convergence theorem*) *Every monotone bounded sequence of real numbers converges to a limit.*

24 See (van Heijenoort 1967) and (Bishop 1967).

25 See (Bridges 1999, p. 96) and (Richman 1990).

26 These results are taken from (Ishihara 2006).

4. CIT: (*The Cantor intersection theorem*).

For MCT (resp. CIT), an algorithm computes the limit (resp. real in the intersection). Next, we list some equivalences of LLPO, the *lesser limited principle of omniscience*. Note that LLPO is an instance of De Morgan's law.

Theorem 12. *In* BISH*, the following are equivalent.*

1. LLPO: $\neg(P \wedge Q) \to \neg P \vee \neg Q \quad (P, Q \in \Sigma_1)$.
2. LLPR: $(\forall x \in \mathbb{R})[\neg(x > 0) \vee \neg(x < 0)]$.
3. NIL: $(\forall x, y \in \mathbb{R})(xy = 0 \to x = 0 \vee y = 0)$.
4. CLO: *For all* $x, y \in \mathbb{R}$ *with* $\neg(x < y)$, $\{x, y\}$ *is a closed set.*
5. IVT: *a version of the intermediate value theorem.*
6. WEI: *a version of the Weierstraß extremum theorem.*

For IVT (resp. WEI), an algorithm computes the interm. value (resp. max.).

It should be noted that any result proved in BISH is compatible[27] with classical, intuitionistic and recursive mathematics.

3.3. Reverse-engineering Reverse Mathematics

In this section, we sketch a translation[28] of results from Constructive Reverse Mathematics to Nonstandard Analysis. We translate[28] Bishop's primitive notion of *algorithm* and *finite procedure* as the notion of Ω-invariance in Nonstandard Analysis. Following Definition 10, the intuitionistic disjunction translates[28] to the following in Nonstandard Analysis.

Definition 13. [Hyperdisjunction] For formulas φ_1 and φ_2, the formula $\varphi_1(n)\ \mathbb{V}\ \varphi_2(n)$ is the statement: *There is an* Ω*-invariant formula* ψ *such that*

$$(\forall n \in \mathbb{N})(\psi(n, \omega) \to \varphi_1(n) \wedge \neg\psi(n, \omega) \to \varphi_2(n). \tag{5}$$

Note that $\varphi_1(n)\,\mathbb{V}\,\varphi_2(n)$ indeed implies $\varphi_1(n) \vee \varphi_2(n)$. Furthermore, given the formula $\varphi_1(n)\,\mathbb{V}\,\varphi_2(n)$, there is an Ω-invariant procedure (provided by $\psi(n, \omega)$) to determine which disjunct of $\varphi_1(n) \vee \varphi_2(n)$ makes it true. Thus, we observe that the meaning of the hyperdisjunction '$\mathbb{V}$' is quite close to its intuitionistic counterpart '$\vee$' from Definition 10.

The other intuitionistic connectives may be translated analogously. The translation of $\to$ (resp. $\neg$) will be denoted $\Rightarrow$ (resp. $\sim$). As for disjunction, the meaning

27 See (Bishop 1967) or (Ishihara 2006).

28 Note that we use the word 'translation' informally: The definition of $\mathbb{V}$ is *inspired* by the intuitionistic disjunction, but that is the only connection.

of the intuitionistic connectives is quite close to that of the hyperconnectives. Furthermore, as suggested by the following theorems, the equivalences from CRM remain valid after the translation. In particular, we have the following theorems, to be compared to Theorems 11 and 12.

Theorem 14. *In Nonstandard Analysis, the following are equivalent.*

1. Π_1-TRANS.
2. $\mathbb{LPO}$: $P \mathbb{V} \sim P \quad (P \in \Sigma_1)$.
3. $\mathbb{LPR}$: $(\forall x \in \mathbb{R})(x > 0 \mathbb{V} \sim (x > 0))$.
4. $\mathbb{MCT}$: (*The monotone convergence theorem*) *Every monotone bounded sequence of real numbers converges to a limit.*
5. $\mathbb{CIT}$: (*The Cantor intersection theorem*).

Analogous to the context of CRM, in $\mathbb{MCT}$ (resp. $\mathbb{CIT}$), the limit (resp. real in the intersection) is computed by an Ω-invariant function.

Theorem 15. *In* $\mathbb{NSA}$*, the following are equivalent.*

1. $\mathbb{LLPO}$: $\sim(P \wedge Q) \Rrightarrow \sim P \mathbb{V} \sim Q \quad (P, Q \in \Sigma_1)$.
2. $\mathbb{LLPR}$: $(\forall x \in \mathbb{R})[\sim(x > 0) \mathbb{V} \sim(x < 0)]$.
3. $\mathbb{NIL}$: $(\forall x, y \in \mathbb{R})(xy = 0 \Rrightarrow x = 0 \mathbb{V} y = 0)$.
4. $\mathbb{CLO}$: *For all* $x, y \in \mathbb{R}$ *with* $\sim(x < y)$, $\{x, y\}$ *is a closed set.*
5. $\mathbb{IVT}$: *a version of the intermediate value theorem.*
6. $\mathbb{WEI}$: *a version of the Weierstraß extremum theorem.*

Analogous to the context of CRM, in $\mathbb{IVT}$ (resp. $\mathbb{WEI}$), the intermediate value (resp. maximum) is computed by an Ω-invariant function.

The previous theorems only constitute an example of a general theme. In particular, it is possible to translate most[29] theorems (and corresponding equivalences) from CRM to Nonstandard Analysis in the same way as above. Comparing Theorems 11 and 12 to Theorems 14 and 15, we conclude that the equivalences observed in CRM remain intact after changing the underlying framework (based on algorithm and intuitionistic logic, by Definition 10) to Nonstandard Analysis (based on Ω-invariance and the hyperconnectives, by Definition 13). Hence, we observe the robustness phenomenon described at the beginning of this section.

In conclusion, we discuss just how far the analogy between Constructive Analysis and Nonstandard Analysis takes us. For instance, on the level of intuition, the

29 See (Sanders 2012) for a list of thirty translated theorems.

formula $\neg(x \leq 0)$ does not imply $x > 0$, as the former expresses that it is impossible that $x \in \mathbb{R}$ is below zero (but might still be *very* close to zero), while the latter expresses that x is bounded away from zero by some rational q we may construct. In Nonstandard Analysis, $\sim(x \leq 0)$ only states that for some (possible infinite) $k \in {}^*\mathbb{N}$, we have $0 < \frac{1}{k} < x$. Hence, $\sim(x \leq 0)$ is consistent with $x \approx 0$, while $x > 0$ has the same interpretation as in BISH. Thus, we observe a correspondence between the latter and Nonstandard Analysis, even on the level of intuitions. A similar conclusion follows from comparing the meaning of IVT and $\mathbb{IVT}$, as is done after Theorem 15.

Secondly, another interesting correspondence is provided by the equivalence between items (1) and (2) in Theorem 12. Indeed, to prove this equivalence, one requires the axiom $\neg(x > 0 \wedge x < 0)$ of the constructive continuum.[30] As it turns out, to establish the equivalence between items (1) and (2) in Theorem 15, the formula $\sim(x > 0 \wedge x < 0)$ is needed in Nonstandard Analysis. Hence, the correspondence between BISH and the latter goes deeper than merely superficial resemblance.

Thirdly, we discuss the above result in the light of the so-called Brouwer-Heyting-Kolmogorov (BHK) interpretation, given by Definition 10. While the equivalences in Theorems 14 and 15 are proved in classical logic, they carry a lot more information. For instance, to show that $\mathbb{LPR}$ implies $\mathbb{LPO}$, a formula $\psi(\vec{x}, n, \omega)$ is defined[29] such that $\psi(\vec{x}, \ulcorner \psi_1 \urcorner, \omega)$ is an Ω-invariant formula which decides between P and $\sim P$ ($P \in \Sigma_1$), for every Ω-invariant formula $\psi_1(\vec{x}, \omega)$ which decides between $x > 0$ and $\sim(x > 0)$. Hence, we do not only have $\mathbb{LPR} \to \mathbb{LPO}$, but also an implication akin to the BHK interpretation, i.e. that an Ω-invariant decision procedure is converted, by an Ω-invariant procedure, to another Ω-invariant decision procedure.

We finish this section with the following remark.

Remark 16 (Reuniting the antipodes)**.** The title of this section refers to a conference with the same name held in 1999 in Venice. Following Bishop's strong criticism[31] of Nonstandard Analysis, this conference was part of a reconciliatory attempts between the communities of Nonstandard Analysis and Constructive Analysis. Little work[32] has indeed taken place in the intersection of these disciplines, but Theorems 14 and 15 can be interpreted as an attempt at *reuniting the antipodes* that are Nonstandard and Constructive Analysis. Nonetheless, it has been noted in the past[33] that Nonstandard Analysis has a constructive dimension.

Acknowledgement: This publication was made possible through the generous support of a grant from the John Templeton Foundation for the project *Philosophical Frontiers in Reverse Mathematics*. I thank the John Templeton Foundation for its continuing support for the Big Questions in science. Please note that

30 See e.g. (Bridges 1999, Axiom set R2).

31 See e.g. (Bishop 1977, p. 208) and (Bishop 1975, p. 513).

32 With the notable exception of Erik Palmgren in e.g. (Palmgren 2001).

33 See (Wattenberg 1988).

the opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation

References

Bishop, E., 1967, *Foundations of Constructive Analysis*. New York: McGraw-Hill Book Co.

Bishop, E., 1975, "The Crisis in Contemporary Mathematics", in: *Proceedings of the American Academy Workshop on the Evolution of Modern Mathematics*, pp. 507-517.

Bishop, E., 1977, "Book Review: Elementary Calculus", in: *Bull. Amer. Math. Soc.* 83, 2, pp. 205-208.

Bridges, D. S., 1999, "Constructive Mathematics: A Foundation for Computable Analysis", in: *Theoret. Comput. Sci.* 219, 1-2, pp. 95-109.

Bridges, D. S. and Vîţă, L. S., 2006, *Techniques of Constructive Analysis*. Universitext, New York: Springer.

Church, A., 1936, "A Note on the Entscheidungsproblem", in: *Journal of Symbolic Logic* 1, pp. 40-41.

Freer, C. E., Ackerman, N. L., and Roy D. M, 2011, "Noncomputable Conditional Distributions", in: *Proceedings of the Twenty-Sixth Annual IEEE Symposium on Logic in Computer Science* (Toronto, Canada, 2011): IEEE Press.

Friedman, H., 1975, "Some Systems of Second Order Arithmetic and Their Use", in: *Proceedings of the International Congress of Mathematicians* (Vancouver, B.C., 1974), vol. 1, Canad. Math. Congress, Montreal, Quebec, pp. 235-242.

Friedman, H., 1976, "Systems of Second Order Arithmetic with Restricted Induction, I & II (Abstracts)", in: *Journal of Symbolic Logic* 41, pp. 557-559.

Hacking, I., 1983, *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge: Cambridge University Press.

Ishihara, H., 2006, "Reverse Mathematics in Bishop's Constructive Mathematics", in: *Philosophia Scientiae* (Cahier Spécial) 6, pp. 43-59.

Keisler, H. J., 2006, "Nonstandard arithmetic and Reverse Mathematics", in: *Bull. Symbolic Logic* 12, 1, pp. 100-125.

Palmgren, E., 2001, "Unifying Constructive and Nonstandard Analysis" (Venice, 1999), in: *Synthese Lib.*, Vol. 306, Dordrecht: Kluwer, pp. 167-183.

Kanovei, V., and Reeken, M., 2004, *Nonstandard Analysis, Axiomatically*. Springer.

Richman, F., 1990, "Intuitionism as a Generalization", in: *Philosophia Math.* 5, pp. 124-128.

Sanders, S., 2011, "ERNA and Friedman's Reverse Mathematics", in: *Journal of Symbolic Logic* 76, pp. 637-664.

Sanders, S., 2012, *On the Notion of Algorithm in Nonstandard Analysis*, Submitted.

Salmon, W. C., 1984, *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Salmon, W. C., 1998, *Causality and Explanation*. Oxford: Oxford University Press.

Simpson, S. G., 2009, *Subsystems of Second Order Arithmetic (Perspectives in Logic)*, 2nd ed., Cambridge: Cambridge University Press.

Soare, R. I., 1987, *Recursively Enumerable Sets and Degrees. Perspectives in Mathematical Logic.* Berlin: Springer-Verlag.

Sommer, R. and Suppes P., 1996. "Finite Models of Elementary Recursive Nonstandard Analysis", in: *Notas de la Sociedad Mathematica de Chile* 15, pp. 73-95.

Sommer, R. and Suppes P., 1997, "Dispensing with the Continuum", in: *Journal of Mathematical Psychology* 41, pp. 3-10.

Turing, A., 1937, "On Computable Numbers, with an Application to the Entscheidungsproblem", in: *Proceedings of the London Mathematical Society* 42, pp. 230-265.

van Heijenoort, J., 1967, *From Frege to Gödel. A Source Book in Mathematical Logic, 1879–1931*. Cambridge (Mass.): Harvard University Press.

Wattenberg, F., 1988, "Nonstandard Analysis and Constructivism?", in: *Studia Logica* 47, 3, pp. 303-309.

Ghent University
Department of Mathematics (S22)
Krijgslaan 281
B-9000, Ghent
Belgium
sasander@cage.ugent.be

Francisco C. Santos and Jorge M. Pacheco

Behavioral Dynamics under Climate Change Dilemmas

Abstract

Preventing global warming is a public good requiring overall cooperation. Contributions will depend on the risk of future losses, which plays a key role in decision-making. Here, we discuss a theoretical model grounded on game theory and large-scale population dynamics. We show how decisions within small groups under high risk and stringent requirements toward success significantly raise the chances of coordinating to save the planet's climate, thus escaping the tragedy of the commons. In addition, our model predicts that, if one takes into consideration that groups of different sizes will be interwoven in complex networks of contacts, the chances for global coordination into an overall cooperating state are further enhanced.

1. Introduction

In a dance that repeats itself cyclically, countries and citizens raise significant expectations every time a new International Environmental Summit is settled. Unfortunately, few solutions have come out of these colossal and flashy meetings, challenging our current understanding and models on decision-making, such that more effective levels of discussion, agreements and coordination become accessible. From Montreal and Kyoto to Copenhagen and Durban summits, it is by now clear how difficult it is to coordinate efforts.[1] Often, individuals, regions or nations opt to be *free riders*, hoping to benefit from the efforts of others while choosing not to make any effort themselves. Cooperation problems faced by humans often share this setting, in which the immediate advantage of free riding drives the population into the Hardin's tragedy of the commons , the ultimate limit of widespread defection.[2]

To address this and other cooperation conundrums ubiquitous at all scales and levels of complexity, the last decades have witnessed the discovery of several core mechanisms responsible to promote and maintain cooperation at different levels of organization. Most of these key principles have been studied within the framework of two-person dilemmas such as the Prisoner's Dilemma, which constitute a powerful metaphor to describe conflicting situations often encountered in the natural and social sciences. Many real-life situations, however, are associated with col-

1 See (Barrett 2005, 2007).

2 See (Hardin 1968).

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_10,

lective action based on joint decisions made by a group often involving more than two individuals. These types of problems are best dealt-with in the framework of N-person dilemmas and Public Goods games, involving a much larger complexity that only recently started to be unveiled. Arguably, the welfare of our planet accounts for the most important and paradigmatic example of a public good: a global good from which everyone profits, whether or not they contribute to maintain it.

One of the most distinctive features of this complex problem, only recently tested and confirmed by means of actual experiments[3], is the role played by the perception of risk that accrues to all actors involved when taking a decision. Indeed, experiments confirm the intuition that the risk of collective failure plays a central role in dealing with climate change. Up to now, the role of risk has remained elusive. Additionally, it is also unclear what is the ideal scale or size of the population engaging in climate summits – whether game participants are world citizens, regions or country leaders – such that the chances of cooperation are maximized. Here we address these two issues in the context of game theory and population dynamics.

The conventional public goods game – the so-called N-person Prisoner's Dilemma – involve a group of N individuals, who can be either Cooperators (C) or Defectors (D). Cs contribute a cost "c" to the public good, whereas Ds refuse to do so. The accumulated contribution is multiplied by an enhancement factor the returns equally shared among all individuals of the group. This implies a collective return which increases linearly with the number of contributors, a situation that contrasts with many real situations in which performing a given task requires the cooperation of a minimum number of individuals of that group.[4] This is the case in international environmental agreements which demand a minimum number of ratifications to come into practice, but examples abound where a minimum number of individuals, which does not necessarily equal the entire group, must simultaneously cooperate before any outcome (or public good) is produced. Furthermore, it is by now clear that the N-person Prisoner's Dilemma fails short to encompass the role of risk, as much as the non-linear nature of most collective phenomena.

Here we address these problems resorting to a simple mathematical model, adopting unusual concepts within political and sustainability science research, such as peer influence and evolutionary game theory. As a result, we encompass several of the key elements stated before regarding the climate change conundrum in a single dynamical model.

In the following we show how small groups under high risk and stringent requirements toward collective success significantly raise the chances of coordinating to save the planet's climate, thus escaping the tragedy of the commons. In other words, global cooperation is dependent on how aware individuals are concerning the risks of collective failure and on the pre-defined premises needed

3 See (Milinski et al. 2008, 2011).

4 E.g., see Alvard, Boesch, Creel, Stander and others.

to accomplish a climate agreement. Moreover, we will show that to achieve stable levels of cooperation, an initial critical mass of cooperators is needed, which will then be seen as role models and foster cooperation.

We will start by presenting the model in Section 2. In Section 3 we discuss the situation in which evolution is deterministic and proceeds in very large populations. In Section 4 we analyze the evolutionary dynamics of the same dilemma in finite populations under errors and behavioral mutations. Finally, in Section 5 we provide a summary and concluding remarks.

2. Model

Let us consider a population of size Z, in which individuals engage in a N-person dilemma, where each individual is able to contribute or not to a common good, i.e., to cooperate or to defect, respectively. Game participants have each an initial endowment b. Cooperators (Cs) contribute a fraction c of their endowment, while defectors (Ds) do not contribute. As previously stated, irrespectively of the scale at which agreements are tried, most demand a minimum number of contributors to come into practice. Hence, whenever parties fail to achieve a previously defined minimum of contributions, they may fail to achieve the goals of such agreement (which can also be understood as the benefit "b"), being this outcome, in the worst possible case, associated with an appalling doomsday scenario.

To encompass this feature in the model we require a minimum collective investment to ensure success: if the group of size N does not contain at least M Cs (or, equivalently, a collective effort of Mcb), all members will lose their remaining endowments with a probability r (the *risk*); otherwise everyone will keep whatever they have. Hence, $M<N$ represents a coordination threshold, necessary to achieve a collective benefit.

As a result, the payoff of a D in a group of size N and k Cs can be written as $\Pi_D(k) = b\,\{\theta(k-M) + (1-r)\,[1-\theta(k-M)]\}$, where $\theta(x)$ is the Heaviside step function ($\theta(x<0)=0$ and $\theta(x\geq 0)=1$. Similarly, the payoff of a C is given by $\Pi_C(k) = \Pi_D(k) - cb$. The risk r is here introduced as a probability, such that with probability *(1-r)* the benefit will be collected independent of the number of contributors in a group.

This collective-risk dilemma represents a simplified version of the game used in the experiments performed by Milinski et al. (2008) on the issue of the mitigation of the effects of climate change, a framework which is by no means the standard approach to deal with International Environmental Agreements and other problems of the same kind. The present formalism has the virtue of depicting black on white the importance of risk and its assessment in dealing with climate change, something that Heal et al. have been conjecturing for quite a while. At

the same time, and unlike Milinski's experiments, our analysis is general and not restricted to a given group size.

Additionally, and unlike most of the canonical treatments, our analysis will not rely on individual or collective rationality. Instead, our model relies on evolutionary game theory combined with one-shot public goods games, in which errors are allowed. In fact, our model includes what we believe are key factors in any real setting, such as bounded rational individual behavior and the importance of risk assessment in meeting the goals defined from the outset.

We assume that individuals tend to copy others whenever these appear to be more successful. Contrary to strategies defined by a contingency plan which, as argued by McGinty before, are unlikely to be maintained for a long time scale, this social learning (or evolutionary) approach allows policies to change as time goes by, and likely these policies will be influenced by the behavior (and achievements) of others, as previously shown in the context of donations to public goods. This also accounts to the fact that agreements may be vulnerable to renegotiation, as individuals may agree on intermediate goals or assess actual and future consequences of their choices to revise their position.

3. Behavioral dynamics in large populations

In the framework of evolutionary game theory, the evolution or social learning dynamics of the fraction x of *C*s (and *1-x* of *D*s) in a large population ($Z \to \infty$) is governed by the gradient of selection *g(x)* associated with the replicator dynamics equation $g(x) \equiv \dot{x} = x(1-x)(f_C - f_D)$, which characterizes the behavioral dynamics of the population, where f_C (f_D) is the fitness of *C*s (*D*s), here associated with the game payoffs. According to the replicator equation, *C*s (*D*s) will increase in the population whenever $g(x)>0$ ($g(x)>0$). If one assumes an unstructured population, where every individual can potentially interact with everyone else, the fitness (or social success) of each individual can be obtained from a random sampling of groups. The latter leads to groups whose composition follows a binomial distribution.[5]

Figure 1 shows the behavior of *g(x)* as a function of the fraction of cooperators (x) for different risk intensities. In the absence of risk ($r = 0.0$), $\dot{x}$ is always negative, leading to the extinction of *C*s ($x = 0$) irrespectively of the initial fraction of cooperators. The presence of risk, in turn, leads to the emergence of two mixed internal *equilibria,* rendering cooperation viable: For finite risk r, both *C*s (below x_L) and *D*s (above x_R) become disadvantageous when rare. Co-existence between *C*s and *D*s becomes stable at a fraction x_R which increases with r. Hence, collective coordination becomes easier to achieve under high-risk, and once the coordination barrier is overcome (x_L), high levels of cooperation will be reached.

5 For details, see (Santos and Pacheco, 2011).

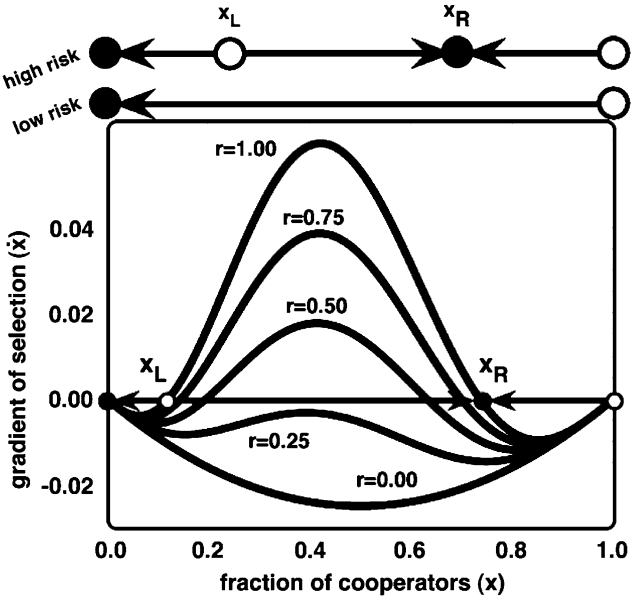


Figure 1. For each fraction of *C*s, if the gradient *g(x)* is positive (negative) the fraction of *C*s will increase (decrease). Increasing risk (*r*) modifies the population dynamics rendering cooperation viable depending on the initial fraction of *C*s (N=6, *M=3* and *c=0.1*).

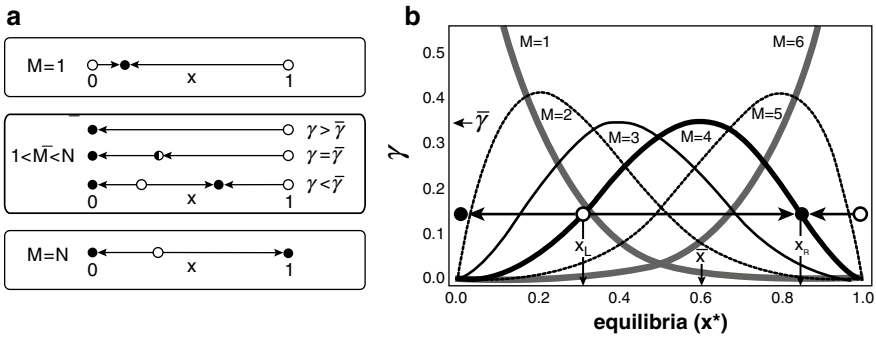


Figure 2. **a)** Classification of all possible dynamical scenarios when evolving an infinitely large population of *C*s and *D*s as a function of g, *M* and *N*. A fraction *x* of an infinitely large population adopts the strategy *C*; the remaining fraction *1-x* adopts *D*. The replicator equation describes the evolution of *x* over time. Solid (open) circles represent stable (unstable) *equilibria* of the evolutionary dynamics; arrows indicate the direction of selection. **b)** Internal roots *x** of *g(x)* for different values of the cost-to-risk ratio $\gamma = c/r$, at fixed group size ($N = 6$) and different coordination thresholds (*M*). For each value of γ one draws a horizontal line; the intersection of this line with each curve gives the value(s) of *x**, defining the inter-

nal *equilibria* of the replicator dynamics. The empty circle represents an unstable fixed point (x_L) and the full circle a stable fixed point (x_R) ($M = 4$ and $\gamma = 0.15$ in example).

The appearance of two internal *equilibria* under risk can be studied analytically.[6] In a nutshell (see also Figure 2a), it can be shown that the location of these *equilibria* can be written down as a function of the *cost-to-risk ratio* γ, defined as $\gamma=c/r$, and coordination threshold M. Scenarios with none, one and two interior fixed points are possible depending if γ is smaller, larger or equal, respectively, to a critical value γ. Hence, the *cost-to-risk ratio* γ plays a central role in dictating the viability of an overall cooperative state:

Intuitively, the smaller the contribution required, the easier it will be to reach such a globally cooperative state. Moreover, the higher the perception of the risk at stake, the more likely individuals react to overcome such a cooperation dilemma.

Figure 2b also shows the role played by the threshold M: for fixed (and low) γ, increasing M will maximize cooperation (increase of x_R) at the expense of making it more difficult to emerge (increase of x_L).

4. Behavioral dynamics in small populations

In reality, however, populations are finite and, in some cases, may be small, as in many collective endeavors, from animal group hunting and warfare, to numerous Human affairs, such as small community collective projects, macroeconomic relations and the famous world summits on climate change, where group and population sizes are comparable and of the order of the hundreds. For such population sizes, stochastic effects play an important role. Stochastic effects are amplified in the presence of errors of different sorts (inducing behavioral *"mutations"*, including errors of imitation). Consequently, they may play an important role in the collective behavior at a population level.

Formally, the population dynamics becomes discrete, whereas the replicator dynamics is no longer valid. Alternatively, we adopt a stochastic process where each individual i imitates the strategy of a randomly selected member of the population j with probability which increases with the fitness difference. Under these circumstances, the behavioral dynamics is best described by a finite population gradient of selection $G(k/Z)$ – defined as the difference of the probabilities to increase and decrease the number k of Cs in the population by one individual – and by the respective stationary distribution of the population, which characterizes the (average) pervasiveness in time of a given fraction of cooperators (k/Z) of the population. Additionally, we consider that, with a small ("mutation") probability, an individual may explore a randomly chosen strategy.

6 For details, see (Santos and Pacheco, 2011).

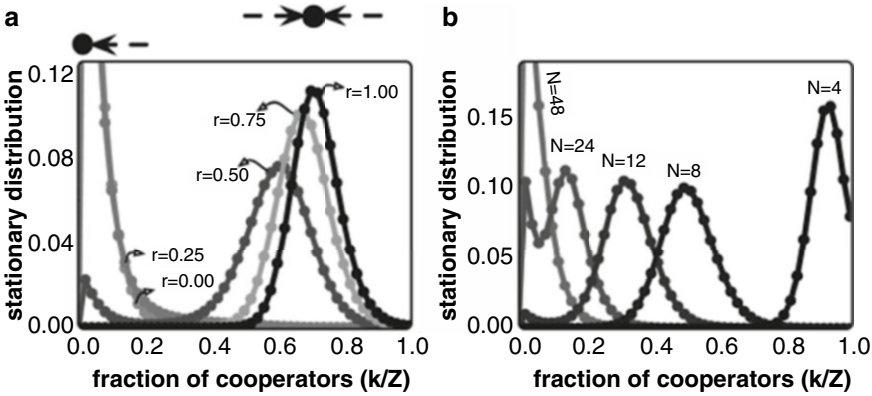


Figure 3. **a)** Stationary distribution describing the prevalence of each fraction (k/Z) of cooperators in finite populations ($Z = 50$ in the presence of mutations and imitation errors). Whenever risk is high, stochastic effects *i*) turn collective cooperation into a pervasive behavior and *ii*) favor the overcome of coordination barriers, rendering cooperation viable, irrespective of the initial configuration ($N = 6$, $M = 3$ and $c = 0.1$). **b)** Stationary distributions for different group sizes and constant threshold $M = 2$. Cooperation will be maximized when risk is high and groups are small (small N), as goal achievement involves stringent requirements.

In Figure 3a we show the stationary distributions for different values of risk, for a population of size $Z = 50$ where $N = 2M = 6$. While the finite population gradient of selection $G(k/Z)$ exhibits a behavior qualitatively similar to $\dot{x}$ in Figure 1, Figure 3a shows that the population spends most of the time in configurations where ***C***s prevail, irrespectively of the initial condition. This is a direct consequence of stochastic effects, which allow the "tunneling" through the coordination barrier associated with x_L, rendering such coordination barrier (x_L) irrelevant and turning cooperation into the prevalent strategy. In short, stochastic effects are able to promote cooperation under collective-risk dilemmas.

Besides perception of risk, group size must also be considered when maximizing the likelihood of reaching overall cooperation, as it defines the scale at which global warming should be tackled. Cooperation for climate control can be achieved at different scales, from regional to global agreements. Hence, even if the problem is certainly global, its solution may be achieved via the combination of several local agreements. So far, attempts have concentrated in a single, global group, although it remains unclear at which scale collective agreements are more easily achieved, as also discussed by Ascheim et al. As shown by the stationary distributions of Figure 3b, cooperation is better dealt with within small groups, even if, for higher M/N values, coordination is harder to attain (see Figure 2).

Figure 3b confirms that with increasing group size cooperation is inhibited, in both scenarios. Given that current policies favor world summits, the present

results suggest a reappraisal of such policies regarding the promotion of public endeavors: instead of world summits, decentralized agreements between smaller groups (small N), possibly focused on region-specific issues, where risk is high and goal achievement involves tough requirements (large relative M), are prone to significantly raise the probability of success in coordinating to tame the planet's climate.

5. Behavioral dynamics in structured populations

The success in self-organizing cooperative behavior within small groups when compared with global dilemmas, naturally begs the question of how these groups should be organized to maximize the chances of cooperation. So far, all groups and individuals have been assumed as identical. Yet, socio-political dynamics is often grounded on a strong diversity in roles and positions. As previously discussed in the context of international agreements, countries are part of intricate networks of overlapping and interrelated alliances or agreements, many of regional nature, involving also geographical neighbors, and others with a global character which transcends geography (see randomly assembled example in Figure 4a). Similarly, diversity in geographical positions, or in social or political configurations, means that some *players* may play a pivotal role in a global outcome, as they may participate in a larger number of 'collective dilemmas' than others.

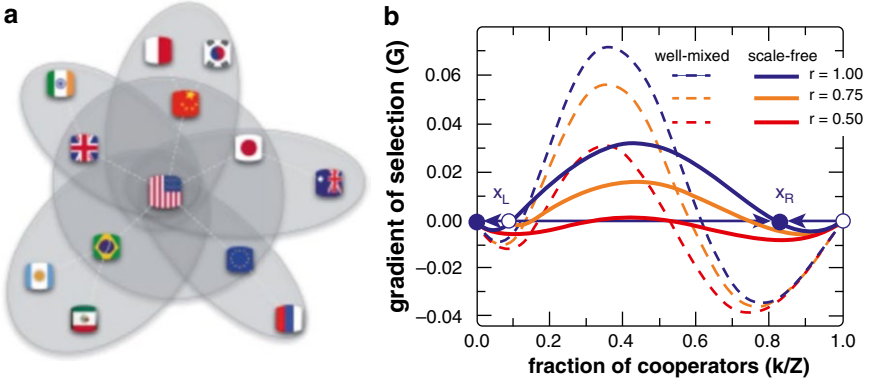


Figure 4. **Evolutionary dynamics in heterogeneous populations. a)** Given an interaction network of size Z and average degree $\langle\zeta\rangle$, where nodes represent individuals, and links represent exchanges or shared goals, *collective-risk dilemmas* may be associated with neighborhoods in this network. As an example, the central individual participates in 6 groups, hence participating in 6 public goods games, each with a given group size. The individual fitness derives from the pay-

off accumulated from all games she/he participates. **a)** Gradients of selection G for a homogeneous (well-mixed) population (dashed lines) and for heterogeneous (scale-free) networks (solid lines), for different values of risk, a population size of $Z = 500$ and an average group size of $<N> = 7 = <\zeta>+1$. In the heterogeneous cases, both size and number of the N-person games each individual participates follow a power-law distribution.

The overall number and size of the dilemmas faced by each individual may be seen as a result of a complex interaction network, where nodes represent individuals, and links represent exchanges, collective investments or shared interests. As exemplified in Figure 4a, each neighborhood of such structure may represent a group with a size given, e.g., by the connectivity of the focal individual. In Figure 4b we show the effect of such heterogeneity or diversity of group sizes in the problem at stake, comparing the finite gradients of selection in a homogeneous setting – taking the well-mixed population as reference – with a heterogeneous case. For the latter, we adopt the ubiquitous power law distribution of connectivities, resulting from a s*cale-free* interaction network and a constant M. This leads to distributions of group sizes and number of games played by each player that also follow a power-law.

As shown in Figure 4b, a heterogeneous contact network changes the location of the internal *equilibria*, without changing either the nature of the effective game or the nature of the internal *equilibria*. However, the impact of such a diversity on the type of game played in each local group is sizeable: in large groups coordination is easier to achieve (M/N is small) but co-existence occurs for a lower fraction of cooperators; in small groups, coordination faces stern requirements (M/N increases) but, once surpassed, most group members will actually cooperate. Whenever the risk of failure is high, introducing group diversity primarily enlarges the stable fraction x_R at equilibrium, also determining a slight increase of the size of the cooperative basin of attraction. Because coordination is easily achieved in large groups, highly connected players at the group centers will acquire a larger fitness. Whenever such *hubs* happen to be occupied by cooperators, they will influence the participants of small groups (the majority) to cooperate, hence enabling small groups to overcome their stringent coordination requirements. Overall, this will act to reduce the average x_L of the population. Once this coordination barrier is surpassed, co-existence will be determined by the small size of the majority of the groups, leading to the dominance of cooperators at x_R.

6. Conclusion

Dealing with environmental sustainability cannot overlook the uncertainty associated with a collective investment. Here we propose a simple form to describe this

problem and study its impact in behavioral evolution, obtaining an unambiguous agreement with recent experiments together with several concrete predictions. We do so in the framework of non-cooperative N-person evolutionary game theory, an unusual theoretical tool within the framework of modeling of political decision-making. We propose a new N-person game where the risk of collective failure is explicitly introduced by means of a simple collective dilemma. Moreover, instead of resorting to complex and rational planning or rules, individuals revise their behavior by peer-influence, creating a complex dynamics akin to many evolutionary systems. This framework allowed us to address the impact of risk in several configurations, from large to small groups, from deterministic towards stochastic behavioral dynamics.

Overall, we have shown how the emerging behavioral dynamics depends heavily on the perception of risk. The impact of risk is enhanced in the presence of small behavioral mutations and errors and whenever global coordination is attempted in a majority of small groups under stringent requirements to meet coactive goals. This result calls for a reassessment of policies towards the promotion of public endeavors: instead of world summits, decentralized agreements between smaller groups (small N), possibly focused on region-specific issues, where risk is high and goal achievement involves tough requirements (large relative M), are prone to significantly raise the probability of success in coordinating to tame the planet's climate. Our model provides a "bottom-up" approach to the problem, in which collective cooperation is easier to achieve in a distributed way, eventually involving regions, cities, *NGO*s and, ultimately, all citizens. Moreover, by promoting regional or sectorial agreements, we are opening the door to the diversity of economic and political structure of all parties, which, as shown can be beneficial to cooperation.

Naturally, we are aware of the many limitations of a bare model such as ours, in which the complexity of Human interactions has been overlooked. From higher levels of information, to non-binary investments, additional layers of realism can be introduced in the model. On the other hand, the simplicity of the dilemma introduced here, makes it generally applicable to other problems of collective cooperative action, which will emerge when the risks for the community are high, something that repeatedly happened throughout Human history, from ancient group hunting to voluntary adoption of public health measures. Similarly, other cooperation mechanisms, known to encourage collective action, may further enlarge the window of opportunity for cooperation to thrive. The existence of collective risks is pervasive in nature, in particular in many dilemmas faced by Humans. Hence, we believe the impact of these results go well beyond decision-making towards global warming.

References

Alvard, M. S., and. Nolin D. A., 2002, "Rousseau's Whale Hunt?: Coordination among Big-Game Hunters", in: *Current Anthropology* 43, 4, pp. 533-559.

Asheim, G. B., Froyn, C. B., Hovi, J., and Menz, F. C., 2006, "Regional Versus Global Cooperation for Climate Control", in: *Journal of Environmental Economics and Management* 51, 1, pp. 93-109.

Barabási, A. L., and Albert, R., 1999, "Emergence of Scaling in Random Networks", in: *Science* 286, 5439 (Oct 15 1999), pp. 509-512.

Barrett, S., 2005, *Environment and Statecraft: The Strategy of Environmental Treaty-Making*. Oxford: Oxford University Press.

Barrett, S., 2007, *Why Cooperate?: The Incentive to Supply Global Public Goods*. Oxford: Oxford University Press.

Black, J., Levi, M. D., and De Meza, D., 1993, "Creating a Good Atmosphere: Minimum Participation for Tackling the 'Greenhouse Effect'", in: *Economica* 60, 239, pp. 281-293.

Boehm, C., 1999, *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge (Mass.): Harvard University Press.

Brewer, N. T., Chapman, G. B., Gibbons, F. X., Gerrard, M., McCaul, K. D., and Weinstein, N. D., 2007, "Meta-Analysis of the Relationship between Risk Perception and Health Behavior: The Example of Vaccination", in: *Health Psychology* 26, 2 (Mar. 2007), pp. 136-145.

Fowler, J. H., and Christakis, N. A., 2010, "Cooperative Behavior Cascades in Human Social Networks", in: *Proceedings of the National Academy of Sciences of the United States of America* 107, 12 (Mar. 23 2010), pp. 5334-5338.

Hardin, G., 1968, "The Tragedy of the Commons", in: *Science* 162, 5364 (Dec. 13 1968), pp. 1243-1248.

Heal, G., 1993, "Formation in International Environmental Agreements", in: C. Carraro (Ed.), *Trade, Innovation, Environment.* Dordrecht: Kluwer.

Heal, G., and Kristrom, B., 2002, "Uncertainty and Climate Change", in: *Environmental and Resource Economics* 22, 1, pp. 3-39.

Kollock, P., 1998, "Social Dilemmas: The Anatomy of Cooperation", in: *Annual Review of Sociolology* 24, pp. 183-214.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., et al., 2009, "Computational Social Science", in: *Science* 323, 5915 (Feb. 6, 2009), pp. 721-723.

McGinty, M., 2010, "International Environmental Agreements as Evolutionary Games", in: *Environmental and Resource Economics* 45, 2, pp. 251-269.

Milinski, M., Röhl, T., and Marotzke, J., 2011, "Cooperative Interaction of Rich and Poor Can Be Catalyzed by Intermediate Climate Targets", in: *Climatic Change*, pp. 1-8.

Milinski, M., Semmann, D., and Krambeck, H. J., "Reputation Helps Solve the 'Tragedy of the Commons'", in: *Nature* 415, 6870 (Jan. 24 2002), pp. 424-426.

Milinski, M., Sommerfeld, R. D., Krambeck, H. J., Reed, F. A., and Marotzke, J., 2008, "The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change", in: *Proceedings of the National Academy of Sciences of the United States of America* 105, 7, pp. 2291–2294.

Ostrom, E., 1990, *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Pacheco, J. M., Santos, F. C., Souza, M. O., and Skyrms, B., 2009, "Evolutionary Dynamics of Collective Action in N-Person Stag Hunt Dilemmas", in: *Proc. Biol. Sci.* 276, 1655 (Jan. 22 2009), pp. 315-321.

Santos, F. C., Santos, M. D., and Pacheco, J. M., 2008, "Social Diversity Promotes the Emergence of Cooperation in Public Goods Games", in: *Nature* 454, 7201 (Jul. 10 2008), pp. 213-216.

Santos, F. C., and Pacheco, J. M., 2011, "Risk of Collective Failure Provides an Escape from the Tragedy of the Commons", in: *Proceedings of the National Academy of Sciences of the United States of America* 108, 26, p. 10421.

Sigmund, K., 2010, *The Calculus of Selfishness*. Princeton: Princeton University Press.

Skyrms, B., 1996, *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, B., 2004, *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Skyrms, B., 2010, *Signals: Evolution, Learning, & Information*. Oxford: Oxford University Press.

Souza, M. O., Pacheco, J. M., and Santos, F. C., 2009, "Evolution of Cooperation under N-Person Snowdrift Games", in: *Journal of Theoretical Biology* 260, 4, pp. 581-588.

Van Segbreck, S., Pacheco, J. M., Lenaerts, T., and Santos, F. C., 2012, "Emergence of Fairness in Repeated Group Interactions", in: *Physical Review Letters* 108, p. 158104.

Francisco C. Santos
GAIPS/INESC-ID, IST Tagusparque
Av. Prof. Dr. Cavaco Silva
2744-016 Porto Salvo
Portugal
franciscocsantos@ist.utl.pt

Jorge M. Pacheco
Departamento de Matemática e Aplicações
Universidade do Minho, Campus de Gualtar
4710-057 Braga
Portugal
pacheco@cii.fc.ul.pt

Sonja Smets

Reasoning about Quantum Actions: A Logician's Perspective

Abstract

In this paper I give an overview of how the work on *quantum dynamic logic for single systems* (as developed in [2]) builds on the concepts of (dynamic) modal logic and incorporates the methodology of *logical dynamics* and *action based reasoning* into its setting. I show in particular how one can start by modeling quantum actions (i.e. measurements and unitary evolutions) in a dynamic logic framework and obtain a setting that improves on the known theorems in traditional quantum logic (stated in the context of *orthomodular lattices*).

1. Introduction

The traditional methods of "static" propositional (and first-order) logic dating back to the first part of the last century are limited with respect to their ability to handle physical systems, especially if we focus on their dynamic, spatial and temporal properties, aspects of uncertainty or probabilistic features. In the meantime several new logical methods have been developed, such as modal logics, in particular propositional dynamic logic (PDL) and temporal logic, dynamic epistemic logics, resource-sensitive logics, game-logics and (in)dependence friendly logics to name just a few. In this paper I follow the dynamic modal logic tradition, which ties in nicely with the work on action logics used in computer science. My aim is to show explicitly how a dynamic modal logic approach can provide the adequate tools to deal with quantum physical systems and moreover, I will point out how this setting provides us with a new methodology to talk about quantum behavior. The methodology fits in line with the dynamic view on logic (as it's practiced by the "Amsterdam school in logic", see e.g. [7, 8]) by focusing, not so much on the 'static' features such as propositions, theories or properties, but on dynamic ones such as: theory change, evaluations, processes, actions, interactions, knowledge updates, communication and observations.

From a more general point of view, the approach adopted in this paper brings together two lines of work: 1) the traditional work on operational quantum logic and 2) a specific information theoretic perspective on quantum systems. As with respect to the first direction, the work on operational quantum logic within the

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_11,

Geneva School on quantum logic originated with [27, 18, 19, 20, 29, 28, 25]. In this work one interprets the logical structure of quantum (and classical) propositions of a physical system by relating it directly to experimental situations. Quantum logic is here not conceived as a "merely" abstract theory (as in [21]) but is provided with an operational dimension which explicitly incorporates features belonging to the realm of "actions and physical dynamics". The second direction refers to the information theoretic part which lines up with the older tradition in computer science of thinking about information systems in a dynamic manner. In this view, a "state" of a system is being identified with the actions that can be (successfully) performed on that state. In theoretical computer science this has given rise to the study of various semantic (classical) notions of "process" (such as e.g. labeled transition systems, automata and coalgebras). Bringing these two lines of work together, I show in the following sections how one can proceed by analogy with the work on labeled transition systems and present a quantum variant of it.

I start in the next section by introducing the necessary background knowledge on labelled transition systems, which is the standard method used in modal logic and in the applications of computer science to represent *processes*. To reason about these processes in Section 3, I go over the standard setting of PDL. In Section 4, I give a quantum interpretation to the language of PDL and show how the setting of quantum transition systems can improve on the known theorems in traditional quantum logic. Note that in this paper no new technical results are being introduced, this paper serves the purpose of highlighting how the classical techniques of modal logics and labeled transitions systems can be adapted and applied to obtain a quantum setting.

2. Labeled Transition Systems

Similar as in [16], I take a *process* to refer to "some object or system whose state changes in time". Note that the logicians' use of process does not necessarily fit in line with the so-called school on *process philosophy*. Broadly viewed, process philosophy relates to the works of Leibniz and Whitehead and is mainly concerned with the ontological nature of processes in the study of metaphysics. One might of course subscribe to the supremacy of processes over other ontological entities, but this is typically not a logician's first concern. Our concern is to reason about processes, in the sense of modeling their behavior.

In modal logic and its applications in computer science, there is a tradition to represent processes by means of *Labeled Transition Systems* (LTS for short), also known as multi-modal Kripke models. A LTS is a structure $(S, \{\stackrel{a}{\rightarrow}\}_{a\in A}, V)$ consisting of a set of states or possible worlds S, a family of binary relations labeled by letters from a given set $a \in A$ and a valuation V assigning truth values to atomic sentences (see e.g. [7]). The set A standardly refers to *actions*, although other interpretations are possible. In a given LTS, defined over a set of actions, the relation $s \stackrel{a}{\rightarrow} t$ indicates that the process can evolve from input state s to

output state t by the execution of action a. As an example, consider the process of getting some money from an ATM modeled as a LTS with basic actions such as "enter your card"; "enter your pin code"; "withdraw 10 euro", etc. Other standard examples of LTS's are e.g. those that encode the process of getting a coffee out of a vending machine, making a zerox-copy or performing a specific calculation on a pocket calculator. In the latter case the arrows are labeled by input-actions such as "$c, +, -, \times, =, 0, \ldots 9$" and states will satisfy strings of input symbols (see e.g. [13]).

It is customary to think of actions as simple, "basic" programs having an input state and an output state. These input and output states represent the internal states of the process. Note that external observers (who push buttons on a pocket calculator) might have no access at all to the internal states that are not visible from the outset. The best picture here is that of a "black box" of which we (the observers/users) only experience its behavior in response to our available actions (see [17, 26]). As explained in [26], the black box picture encodes the difference between an LTS and a finite state automaton. In a finite state automaton one first has to provide an input list and then one lets the automaton run to decide if it accepts or rejects the input. Contrary to an LTS, in an automaton one does not see immediately whether each action (that provides an input-item) is rejected or not, one has to wait until the automaton eventually stops running. Further, LTS's can have an infinite amount of states and hence they differ in an obvious way from finite state automata.

Several types of processes can be represented by the formalism of LTS. Nondeterministic processes can be captured by using branching relations to represent "arbitrary choice". Similarly, the LTS-formalism can capture the concatenation and iteration of processes by using the composition of transition relations.

Stochastic processes are essentially probabilistic and can be represented by probabilistic versions of labeled transition systems. In the case of a *discrete* state space, the study of probabilistic transition systems was initiated by Larsen and Skou in [23]. They define a *probabilistic transition system* $(S, \{\mu_{s,a}\})$ as a structure consisting of a set of states S and a family of probability distributions $\mu_{s,a}$, one for each action $a \in A$ and each input-state $s \in S$. Here, $\mu_{s,a} : S \mapsto [0, 1]$ gives the possible next states (and their probabilities) after action a is performed on input-state s. Informally, $\mu_{s,a}(s') = x$ says that action a can be performed in state s and with probability x reaches the state s' afterwards [23]. Note that the probabilities have to add up: $\sum_{s'} \mu_{s,a}(s') = 1$.

Larsen and Skou's investigation was first extended to the case of *continuous* state spaces in [10]. As explained in [11], this means that "we cannot ask for the transition probability [from an input-state] to any [specific output-state, or some arbitrary] set of states - we need to restrict ourselves to *measurable* sets". In such a setting, one can model the complex continuous real-time stochastic systems such as the flight management system of an aircraft or the Brownian motion of some molecules [12].

I briefly note here the existence of a general abstract mathematical framework encompassing and unifying all the above-mentioned types of processes and many others: the *theory of coalgebras* (see e.g. [17, 22]). Coalgebra is a rather new domain of research, drawing mainly upon the mathematical language of Category Theory. A coalgebra, in its most rudimentary form, consists of a state space S endowed with a transition map $S \to F(S)$, where F is a functor. By varying the functor, one can accommodate many possible notions of processes: transition systems, deterministic systems, discrete probabilistic systems, continuous stochastic systems etc. For the purpose of this paper, I will not go further into the general framework of coalgebras, restricting myself to the simplest example above (non-probabilistic labeled transition systems). But it is important to stress that, from a general coalgebraic perspective, the above discussion can be extended to other types of processes.

3. Propositional Dynamic Logic

One of the logical systems that provides an axiomatic proof theory to reason about the actions in a LTS is Propositional Dynamic Logic (PDL). PDL and its fragment the Hoare Logic (see e.g. [15]) have been mainly used in the context of *program verification in computer science*, i.e. when verifying that a given (classical) action or program meets a required specification. In its syntax, PDL uses dynamic formulas to express these actions or programs. Besides the basic actions that were introduced in the previous section, PDL also considers some special kind of actions, called *"tests"*. Each classical property $P \in \mathcal{P}(S)$ gives rise to a "test" denoted as $P?$. Hence, the actions of PDL could be classified in two types: tests $P?$ and basic actions A. Semantically this means that I slightly generalize the above given semantic setting to incorporate the two types of actions as follows:

A *dynamic frame* is a structure $\mathcal{F} = (S, \{\overset{P?}{\to}\}_{P\in\mathcal{L}}, \{\overset{a}{\to}\}_{a\in A})$, consisting of a set S of *states*; a family of binary "transition" relations $\overset{P?}{\to} \subseteq S \times S$, which are labeled by "test" actions $P?$; a family of binary "transition" relations $\overset{a}{\to} \subseteq S \times S$, labeled by basic "actions" $a \in A$. Note that the labels for the tests come from a given family $\mathcal{L} \subseteq \mathcal{P}(S)$ of subsets $P \subseteq S$, which are called *testable properties*.

As noted in [1, 2, 3], *Kripke frames for standard* PDL are a special case of dynamic frames, namely those in which one takes $\mathcal{L} =: \mathcal{P}(S)$, and the transition relation for a test to be given by $s \overset{P}{\to} t$ iff $s = t \in P$. Semantically this is encodes as the *diagonal* $\{(w, w) : w \in P\}$ of the set P. As noted in [2], intuitively $P?$ can be thought of as a "purely epistemic" action by a (external) observer who "tests" property P, without affecting the state of the system. The transitions $\overset{a}{\to}$ are binary relations on S.

The logical language of standard (star-free) PDL consists of two levels: a level of propositional *sentences* φ (expressing properties) and a level of *programs or actions* π which are defined by mutual induction:

$$\begin{array}{lllll} \varphi & ::= & p| & \neg\varphi| & \varphi \wedge \varphi| \quad [\pi]\varphi \\ \pi & ::= & a| & \varphi?| & \pi \cup \pi| \quad \pi;\pi \end{array}$$

Here I take $p \in \Omega$ and Ω to be a given set of basic (elementary) propositions. The set of basic action labels A is given with $a \in A$. I use $\neg$ to denote *classical negation* and $\wedge$ for *classical conjunction*. The modal operators $[\pi]$ are labeled by actions π and in this I allow for complex action or program constructions such as the non-deterministic choice of actions $\pi \cup \pi$ and relational composition $\pi;\pi$. I use labeled modal operators to build a particular type of formulas $[\pi]\varphi$, which construct a new formula from a given program π and formula φ. Here $[\pi]\varphi$ is used to express weakest preconditions, which means that if program π would be performed on the current state of the system then the output state will necessarily satisfy φ. The PDL test $\varphi?$ denotes the action of testing for φ in the way as is defined in the above semantics for standard PDL. Hence the test action $\varphi?$ is successful if and only if φ is true and testing for φ leaves the state of the system unchanged, in all other cases the test fails. In line with [6], we note that all these complex program constructors make PDL particularly well fit for the task of program verification as it becomes easy in this setting to express programming constructs such as "if then else" or "do while"-loops (see [15]). In a way this indicates the importance of lifting this setting to a quantum framework, precisely because of the contributions it can offer to the work on quantum program (or quantum protocol) verification (as in [1, 3]).

4. Dynamic Quantum Logic

In this section, I show how the ideas presented in the previous sections can be extended to a quantum framework. I don't present new technical results here but provide an overview of the main ideas of Dynamic Quantum Logic as presented in a series of papers [2, 1, 3, 4, 5, 6, 30].

In [2] it was first shown how Hilbert spaces can be structured as non-classical relational models. These models are a quantum version of the LTS's introduced above. I call a *Quantum Transition Systems* (or QTS) a dynamic frame $(S, \{\overset{P?}{\rightarrow}\}_{P\in\mathcal{L}}, \{\overset{a}{\rightarrow}\}_{a\in A})$ satisfying a set of ten abstract semantic conditions. In this case the states in S are meant to represent the *possible states of a quantum physical system* and the transition relations describe the changes of state induced by the *possible actions* that may be performed on that quantum system. As before I use the notation $\mathcal{L}$ to denote the set of testable properties. Any such QTS can be equipped with a so-called *measurement relation*, which allows for the existential quantification over tests as follows: $s \rightarrow t$ iff $s \overset{P?}{\rightarrow} t$ for some $P \in \mathcal{L}$. The negation of the measurement relation gives rise to an *orthogonality relation*, so I write $s \perp t$ iff $s \not\rightarrow t$. For any set $P \subseteq S$, I write $t \perp P$ iff $t \perp s$ *for all* $s \in P$ and the *orthogonal* (or *orthocomplement*) of the set P is defined as follows: $\sim P := \{t \in S : t \perp P\}$. The *biorthogonal closure of a set* P is given by the set $\sim\sim P = \sim(\sim P)$.

In the following list of semantic frame conditions in a QTS, I take the variables P, Q to range over testable properties in $\mathcal{L}$, the variables s, t, s', t', v, w range over states in S and a ranges over basic actions (which are also called "unitary evolutions") [2]:

Frame Conditions

1. Closure under *arbitrary conjunctions*: if $\mathcal{L}' \subseteq \mathcal{L}$ then $\bigcap \mathcal{L}' \in \mathcal{L}$

2. *Atomicity*. States are testable, i.e. $\{s\} \in \mathcal{L}$.
   This is equivalent to requiring that "states can be distinguished by tests", i.e.
   if $s \neq t$ then $\exists P \in \mathcal{L} : \ s \perp P, \ t \not\perp P$

3. *Adequacy*. Testing a true property does not change the state:
   if $s \in P$ then $s \xrightarrow{P?} s$

4. *Repeatability*. Any property holds after it has been successfully tested:
   if $s \xrightarrow{P?} t$ then $t \in P$

5. *"Covering Law"*. If $s \xrightarrow{P} w \neq t \in P$, then there exists some $v \in P$ such that $t \rightarrow v \not\rightarrow s$.

6. *Self-Adjointness Axiom*: if $s \xrightarrow{P?} w \rightarrow t$ then there exists some element $v \in S$ such that $t \xrightarrow{P?} v \rightarrow s$

7. *Proper Superposition Axiom*. Every two states of a quantum system can be properly superposed into a new state: $\forall s, t \in S \ \exists w \in S \ s \rightarrow w \rightarrow t$

8. *Reversibility and Totality Axioms*. Basic unitary evolutions are total bijective functions: $\forall t \in S \ \exists! s \ s \xrightarrow{a} t$ and $\forall s \in S \ \exists! t \ s \xrightarrow{a} t$

9. *Orthogonality Preservation*. Basic unitary evolutions preserve (non) orthogonality: Let $s, t, s', t' \in S$ be such that $s \xrightarrow{a} s'$ and $t \xrightarrow{a} t'$.
   Then: $s \rightarrow t$ iff $s' \rightarrow t'$.

10. *Mayet's Condition: Orthogonal Fixed Points*. There exists some unitary evolution $a \in A$ and some property $P \in \mathcal{L}$,such that a maps P into a proper subset of itself; and moreover the set of fixed-point states of a has dimension ≥ 2. In other words:
    $\exists a \in A \exists P \in \mathcal{L} \exists t, w \in S \forall s \in \sim\sim \{t, w\} : \ a(P) \subseteq P, a(P) \neq P, t \perp w, a(s) = s$.

As shown in [2], these 10 conditions imply that $\mathcal{L}$, with set-inclusion as partial order, forms an *orthomodular lattice* of infinite height satisfying all the necessary conditions for the representation theorem of Piron, Solèr and Mayet to hold (see [28, 24, 31]). To understand this result, let us first call a *concrete* QTS a QTS

which is given by an infinite-dimensional Hilbert space H. In the concrete case, the "states" in S are taken to the *one-dimensional closed linear subspaces of* H, $\mathcal{L}$ is then given by the family of *closed linear subspaces of* H and the relations that are labeled by testable properties $\xrightarrow{P?}$ will correspond to (successful) quantum tests (given by the projectors onto the closed linear subspace corresponding to property P). The relations $\xrightarrow{a}$ correspond to linear maps (expressing the so-called "quantum gates") a on H. The important result for this setting, proved in [2], shows an "Abstract Soundness and Completeness" theorem for the Hilbert-space semantics. In particular, the results in [2] show how every (abstract) QTS can be canonically embedded in the concrete QTS associated to an infinite-dimensional Hilbert space, i.e. every concrete QTS is a QTS and every QTS is isomorphic to a concrete QTS.

We argued in [2, 4, 5, 6] for the importance of these results. By moving to the QTS setting it is possible to solve some of the main problems posed in the traditional quantum logic work on orthoframes, such as the problem that orthomodularity could not be captured by a first-order frame condition (as shown in [14]). In contrast, in a QTS this problem is solved: orthomodularity now does correspond to a first-order frame condition and receives a natural dynamic interpretation. In a similar fashion we refrased the "Mayet condition", which previously could only be stated using the second-order notion of a lattice isomorphism. The "Mayet condition" has now been "internalized" in the setting via the use of quantum actions. Hence from a logical perspective, the QTS formalism yields an improvement of the traditional quantum logic setting.

The QTS structures provide us with the models for a propositional logical system that is different but still close to traditional PDL. The logic is called the *Logic of Quantum Actions* (LQA) in [2, 4, 5, 6] and has the same syntactic language as (star-free) PDL. Let us restrict our quantum setting to the language without classical negation $\neg$ in this paper. This (star-free and classical negation-free) language of PDL can be interpreted in a QTS. All the actions are now interpreted as quantum actions, in particular the test operation will correspond to a quantum test and a basic action is interpreted as a quantum gate. The complex program expressions can now be interpreted as quantum programs.[1]

Note that traditional orthomodular quantum logic (in the tradition of work by [9]) can be re-interpreted inside LQA. This can be done by defining the orthocomplement of a property as the impossibility of a successful test, i.e. $\sim \varphi := [\varphi?]\bot$. Note that the operation of "quantum join" is definable via de Morgan law as $\varphi \sqcup \psi :=\sim (\sim \varphi \wedge \sim \psi)$ and the traditional "quantum implication" (or so-called *Sasaki hook*) is given by the weakest precondition $\varphi \rightarrow \psi := [\varphi?]\psi$. This re-interpretation provides us with a dynamic and operational characterization of all the non-classical connectives of traditional quantum logic.

1 The setting can be extended with a classical negation, which then means that not all "sets of states" $P \subseteq S$ will correspond to "quantum testable properties". In [4, 5] we showed how this enriches the setting and gives us more expressive power than traditional Quantum Logic.

Acknowledgement: The research presented here has been made possible by VIDI grant 639.072.904 of the Netherlands Organization for Scientic Research (NWO).

References

[1] Baltag A. and Smets S., 2004, "The Logic of Quantum Programs", in: P. Selinger (Ed.) *Proceedings of the 2nd International Workshop on Quantum Programming Languages (QPL2004), TUCS General Publication*, 33, Turku Center for Computer Science, pp. 39-56.

[2] Baltag A. and Smets S., 2005, "Complete Axiomatizations of Quantum Actions", in: *International Journal of Theoretical Physics*, 44, 12, pp. 2267-2282.

[3] Baltag A. and Smets S., 2006, "LQP: The Dynamic Logic of Quantum Information", in: *Mathematical Structures in Computer Science*, Special Issue on Quantum Programming Languages, 16, 3, pp. 491-525.

[4] Baltag A. and Smets S., 2008, "A Dynamic-Logical Perspective on Quantum Behavior", in: L. Horsten and I. Douven (Eds.), *Special Issue: Applied Logic in the Philosophy of Science*, *Studia Logica*, 89, pp. 185-209.

[5] Baltag A. and Smets S., 2011, . "Quantum Logic as a Dynamic Logic", in: T. Kuipers, J. van Benthem and H. Visser (Eds.), *Synthese*, 179, 2, pp. 285-306.

[6] Baltag A. and Smets S., 2011, "The Dynamic Turn in Quantum Logic", in: *Synthese*, Online First at Springer.

[7] van Benthem J., 1996, *Exploring Logical Dynamics*, Stanford, CA: CSLI Publications.

[8] van Benthem J., 2011, *Logical Dynamics of Information and Interaction*, Cambridge: Cambridge University Press.

[9] Birkhoff G., and von Neumann J., 1936, "The Logic of Quantum Mechanics", in: *Annals of Mathematics*, 37, pp. 823-843, reprinted in: C. A. Hooker (Ed.), 1975, *The Logico-algebraic Approach to Quantum Mechanics*, vol. 1, Dordrecht: D. Reidel Publishing Company, pp .1-26.

[10] Blute R., Desharnais J., Edalat A. and Panangaden P., 1997, "Bisimulation for labelled Markov Processes", in: *Proceedings of the Twelfth IEEE Symposium on Logic in Computer Science, Warsaw, Poland*.

[11] Danos V., Desharnais J., Laviolette F., and Panangaden P., 2006, "Bisimulation and Cocongruence for Probabilistic Systems", in: *Information and Computation* 204, 4, pp. 503-523.

[12] Desharnais J., Gupta V., Jagadeesan R. and Panangaden P., 2003, "Approximating Labeled Markov Processes", in: *Information and Computation* 184, 1, pp. 160-200.

[13] Fokkink W., 2000, *Introduction to Process Algebra*. Germany: Springer.

[14] Goldblatt R., 1984, "Orthomodularity Is Not Elementary", in: *The Journal of Symbolic Logic* 49, pp. 401-404.

[15] Harel D., Kozen D. and Tiuryn J., 2000, *Dynamic Logic*. Cambridge (Mass.): The MIT Press.

[16] Hartmann S., 1996, "The World as a Process: Simulations in the Natural and Social Sciences" in: R. Hegselmann et al. (Eds.), *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, Theory and Decision Library, Dordrecht: Kluwer, pp. 77-100.

[17] Jacobs B., n.d., *Introduction to Coalgebra. Towards Mathematics of States and Observations.* Book in preparation, on-line at http://www.cs.ru.nl/ bart/PAPERS/index.html

[18] Jauch, J. M., 1968, *Foundations of Quantum Mechanics*, Addison-Wesley, Reading, Massachusetts.

[19] Jauch, J. M. and Piron C., 1969, "On the Structure of Quantal Proposition Systems", in: *Helvetica Physica Acta* 42, pp. 842-848.

[20] Jauch, J. M., and Piron C., 1970, "What is 'Quantum-Logic'?" in: P. G. O. Freund, C. J. Goebel, and Y. Nambu (Eds.) *Quanta*. Chicago: The University of Chicago Press.

[21] Kalmbach G., 1983, *Orthomodular Lattices*. London–New York: Academic Press.

[22] Kurtz A., 2000, *Logics for Coalgebras and Applications to Computer Science*, PhD-thesis, München, Germany.

[23] Larsen K. G. and Skou A., 1991, "Bisimulation Through Probabilistic Testing", in: *Information and Computation* 94, pp. 1-28.

[24] Mayet R., 1998, "Some Characterizations of the Underlying Division Ring of a Hilbert Lattice by Automorphisms", in: *International Journal of Theoretical Physics* 37, 1, pp. 109-114.

[25] Moore D. J., 1999, "On State Spaces and Property Lattices", in: *Studies in History and Philosophy of Modern Physics* 30, pp. 61-83.

[26] Panangaden P., 2006, "Notes on Labelled Transition Systems and Bisimulation", course-notes, October.

[27] Piron C., 1964, *Axiomatique quantique (PhD-Thesis)*, *Helvetica Physica Acta*, 37, pp. 439-468; English Translation by Cole M.: *Quantum Axiomatics* RB4 Technical memo 107/106/104, GPO Engineering Department, London.

[28] Piron C., 1976, *Foundations of Quantum Physics*. W.A. Benjamin Inc., Massachusetts.

[29] Piron C., 1990 *Mécanique quantique. Bases et applications*, Presses polytechniques et universitaires romandes, Lausanne (Second corrected ed. 1998).

[30] Smets S., 2010, "Logic and Quantum Physics", in: A. Gupta and J. van Benthem (Eds.), *Journal of the Indian Council of Philosophical Research Special Issue* XXVII, 2.

[31] Solèr M. P., 1995, "Characterization of Hilbert Spaces by Orthomodular Spaces", in: *Communications in Algebra* 23, 1, pp. 219-243.

Institute for Logic, Language and Computation
University of Amsterdam
P.O. Box 94242
1090 GE, Amsterdam
The Netherlands
S.J.L.Smets@uva.nl

Leszek Wroński

Branching Space-Times and Parallel Processing[1]

Abstract

There is a remarkable similarity between some mathematical objects used in the Branching Space-Times framework and those appearing in computer science in the fields of event structures for concurrent processing and Chu spaces. This paper introduces the similarities and formulates a few open questions for further research, hoping that both BST theorists and computer scientists can benefit from the project.

1. Introduction

The goal of this short paper is to put forward a few open questions regarding the connections between two areas, one mainly of interest to philosophers, the other to computer scientists: the theory of Branching Space-Times (BST) and the field of modelling parallel processing. The hope is that establishing these connections will eventually help to solve some fundamental technical difficulties of the BST approach, while allowing some types of structures from the realm of computer science to have a spatiotemporal representation.

Why should a theory which, judging by its name, concerns branching space-times, be in any way connected to parallel processing? Consider first the well known theory of Branching Time (BT): any BT structure can be viewed as modelling the way a certain indeterministic process could go. It would seem that a theory which allows modelling of *bundles* of (possibly) indeterministic processes is just a step away, requiring only the modification of the representation of maximal possible courses of events: they should no longer be linear, but should have a spatial dimension. This, however, would still not be enough to capture the idea of independent choices (or indeterministic events), and thus the relationship between BT and BST is a bit more complicated. In the next section we introduce the two approaches and state the first open problem about BST. In the two sections that follow we sketch two approaches to parallel processing in computer science: that

1 This paper stems from a joint project with Thomas Müller (Universiteit Utrecht), who told me of the idea, triggered by a remark by Hu Liu, of connecting the Branching Space-Times theory to the approaches to parallel processing found in computer science.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_12,

of "event structures" (ES) and that of Chu spaces. The results we have in mind concern the methods of generating structures of a given type given a structure of another type so that some important "structural" infomation is preserved; for example, how to construct a Chu space given a BST model and *vice versa*. A method for generating an event structure on the basis of a Chu Space is sketched at the end of Section 4.

2. Branching Time and Branching Space-Times

2.1 Branching Time (BT) structures

Definition 1 A *BT structure* is a pair $\langle W, \leqslant \rangle$ such that:

- $W \neq \emptyset$;
- $\leqslant$ is a partial order on W;
- $\leqslant$ is backward-linear.

A *history* is a maximal chain in $\langle W, \leqslant \rangle$.

One intended interpretation of the above considers W to be the set of possible events understood as time-slices through the whole universe and $\leqslant$ to be the "earlier-possibly later" relation. Each history represents one complete way the world could unfold.

In philosophy, BT has been widely used, especially in discussions of agency and future contingents. Unfortunately, for some goals the approach is unwieldy: in any history any two events are ordered. This is not convenient if one has in mind portraying independent choices of two agents or modelling experiments which take part in spatiotemporally separated regions.

To overcome this difficulty, a natural first step is to make events "smaller" – they should represent the action in bounded spatiotemporal regions or even, ideally, point events.[2] If so, then histories can no longer be chains. The guiding idea of the BST approach is that histories should represent space-times.

2 For the discussion of point events see the founding paper for the BST theory: Nuel Belnap, 1992, "Branching Space-Time", in: *Synthese* 92, 3, pp. 385-434.

2.2 Branching Space-Times structures

In the words of its creator, the goal of the BST theory is to "combine *relativity* and *indeterminism* in a rigorous theory".[3] The indeterministic aspect is carried over from the BT approach, and the relativistic aspect is to be achieved by re-imagining the notion of history. However, BST is not a straightforward generalization of BT: it will become evident after the definition of a BST structure is given that not all BT structures are BST structures.

The definition of a BST structure is significantly more complicated than that of a BT structure; it refers to the "new" notion of history and the notion of a choice point, which we will now define.

Definition 2 A *BST history* is a maximal upward directed set. (A set is upward directed iff for any two its elements e_1, e_2 it contains an element e such that $e_1 \leqslant e$ and $e_2 \leqslant e$.)

A *choice point* between two histories h_1 and h_2 is a point e maximal in the intersection $h_1 \cap h_2$.

We say that points e and f are *space-time related* (SLR) if there is a history h such that $e, f \in h$ but neither $e \leqslant f$ nor $f \leqslant e$.

In BST, we have two types of "forward branching": modal (think of a real choice, an event having two possible futures) and non-modal (e.g. emission of particles from a source). In contrast, there is *no modal backward branching*: every event has a fixed past. This is because events are to be thought of as *tokens*, not *types*.

Definition 3 A *BST structure* is a tuple $\langle W, \leqslant \rangle$, where:

- $W \neq \emptyset$;
- $\leqslant$ is a partial order on W;
- $\leqslant$ is dense in W;
- W has no maximal elements w.r.t. $\leqslant$;
- every lower bounded chain in W has an infimum in W;
- every upper bounded chain in W has a supremum in every history it is a subset of;
- (Prior choice principle ('PCP')) for any lower bounded chain $O \subseteq h_1 - h_2$ there exists a choice point $e \in W$ for h_1 and h_2 such that $\forall e' \in O\ e < e'$.

We do not have space here to present a detailed motivation of the conditions; they mostly stem from the two ideas of histories representing space-times (for which the

3 *Ibid.*, p. 385.

conditions are still not enough; see below) and events being understood as token entities.

In BST W is interpreted as containing all possible *point* events (ideally, each event is located in a single space-time point). $\leqslant$ is to be read as the ordering of possible causal influence, frequently interpreted as the light cone ordering: $e \leqslant f$ iff f is in the future light cone of e.

There have been two main areas of applying BST. Some researchers used the approach to model the various experiments connected with the Bell theorem. Perhaps the non-probabilistic GHZ setup proved to be the most tractable by means of BST, but the probabilistic setups have also been the topic of discussion among BST theorists.[4] The second area to which BST has been recently applied is that of agency.[5]

We will soon define the so called "transition structures" of BST structures; a transition structure will be a "skeleton" of the given BST structure, containing all the important information about branching. The current project aims to explore the remarkable similarity of these transition structures to some structures found in computer science – most notably the "event structures for concurrent processes" and Chu spaces. In the process we hope to answer some open problems about BST and provide a new, spatiotemporal reading to the aforementioned structures.

First, though, a disclaimer is in order. We said above that in BST histories are to represent space-times. The definition of a BST structure is, unfortunately, not enough for this. There are BST structures which cannot be provided with a useful notion of a space-time point.[6] Still, there is a class of BST structures, called "Minkowskian Branching Structures" (MBSs), in which all histories are isomorphic to the Minkowski space-time.[7] Since relativistic aspects are not relevant for the task at hand, we will assume that all considered BST structures are MBSs and thus we can think of histories as of copies of the Minkowski space-time.

For example, the following picture represents a BST structure with four histories and two binary choice points:

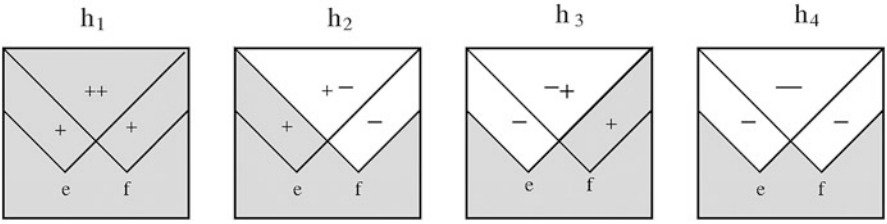
---

4 See e.g. Nuel Belnap and László Szabó, 1992, "Branching Space-Time Analysis of the GHZ Theorem", in: *Foundations of Physics* 26, 8, pp. 989-1002; and Tomasz Placek, 2010, "On Propensity-Frequentist Models for Stochastic Phenomena with Applications to Bell's Theorem", in: Tadeusz Czarnecki, Katarzyna Kijania-Placek, Olga Poller and Jan Woleński (Eds.), *The Analytical Way*, London: College Publications, pp. 105-140.

5 See e.g. Nuel Belnap, 2011, "Prolegomenon to Norms in Branching Space-Times", in: *Journal of Applied Logic* 9, pp. 83-94.

6 See Thomas Müller, 2005, "Probability Theory and Causation: A Branching Space-Times Analysis", in: *British Journal for the Philosophy of Science* 56, 3, pp. 487-520.

7 See Thomas Müller, 2002, "Branching Space-Time, Modal Logic and the Counterfactual Conditional", in: Tomasz Placek and Jeremy Butterfield (Eds.), *Non-locality and Modality*, Dordrecht: Kluwer, pp. 273-291 and Leszek Wroński and Tomasz Placek, 2009, "On Minkowskian Branching Structures", in: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 40, 3, pp. 251-258.



where the shaded sections denote the regions of intersection of the particular history with h_1.

2.3 Where the action is: transitions

BST structures are "big" in the sense that they encompass whole space-times; it would seem however that in some cases – like the simple 4-history structure above – it should be possible to distill the information about "what really happens" in the model and store it in some discrete format. According to the way of thinking about what happens in BST structures established in the literature any action goes on at a choice point, in which an "immediate possibility" is chosen. What happens in a BST structure are "transitions". In the following assume $Hist$ to be the set of all histories in the model and $H_{(e)}$ to be the set of histories containing the point e.

Definition 4 We say that histories h_1 and h_2 *do not divide* at e ($h_1 \equiv_e h_2$) iff $\exists_{e^*>e}\ e^* \in h_1 \cap h_2$ (the two histories share a point above e).

Since $\equiv_e$ is an equivalence relation on $H_{(e)}$, we can, for any e, partition the $H_{(e)}$ into the equivalence classes of $\equiv_e$, to be thought of as "immediate possibilities (open) at e".

Definition 5 A *transition* is a pair

$$\langle \text{ a choice point } e, \text{ an immediate possibility open at } e \ \rangle.$$

We will assume the usual practice of denoting transitions using arrows. For example, in the picture above, we can label the two immediate possibilities open at the binary choice point e as "$+_e$" and "$-_e$" and consider two transitions $e \rightarrowtail +_e$ and $e \rightarrowtail -_e$.

The set $TR(OW)$ of all transitions in a BST structure OW[8] can be given a natural partial order $\leqslant_T$ by taking the reflexive closure of $<_T$, defined as follows, for $t_i = e_i \rightarrowtail H_i$:

$$t_i <_T t_j \text{ iff } e_i < e_j \text{ and } H_{(e_j)} \subseteq H_i.$$

8 This is an abbreviation dating back to the 1992 paper by Belnap (*op. cit.*), denoting "Our World".

One could expect that, given a BST structure, its transition structure together with the information about the location of branching should be enough to store all the "relevant" information about the initial structure in a discrete format; relevant in the sense that (a structure isomorphic to) the initial structure should be recoverable from the discrete data. It turns out that this is not true: as we will see in the next subsection, there are different BST structures which share the same transition structure.

2.4 Consistency and modal unsaturation

We will call a set of transitions "consistent" if, colloquially, they can all happen together, or – in other words – there is a history in which all of them are realised. Formally, a set of transitions $\{t_i \rightarrowtail H_i \mid i \in I\}$ is *consistent* if $\exists h \in Hist\ h \in \bigcap_{i \in I} H_i$; it is *inconsistent* otherwise. Two different transitions from the same event e are called *blatantly inconsistent*. Histories correspond to maximal consistent sets of transitions.

We now come to an interesting feature of some BST structures (in fact, it may be *the* feature which makes them useful for modelling quantum experiments): that of *modal unsaturation*. (Usually called "funny business" in the literature; we will sometimes also use this term.) The essence is this: it might happen that some "combinatorically possible" history is missing from the structure – for example we can have two SLR binary choice points e and f, four transitions $e \rightarrowtail +_e$, $e \rightarrowtail -_e$, $f \rightarrowtail +_f$ and $f \rightarrowtail -_f$, such that the only two pairs of consistent transitions are $\{e \rightarrowtail +_e, f \rightarrowtail +_f\}$ and $\{e \rightarrowtail -_e, f \rightarrowtail -_f\}$! In other words, if you deleted the histories h_2 and h_3 from the structure depicted on p. 139, you would end up with a perfectly fine BST structure, with *exactly the same* transition structure.

To sum up: the goal is to give a discrete representation of a given BST structure OW using the structure of its transitions, $TR(OW)$, and the information about the location of branching. An isomorphic BST structure should be recoverable from the discrete representation. This goal has been achieved by Thomas Müller in 2010,[9] but only for modally saturated BST structures. It would seem that the discrete format for BST structure representation should simply contain another "module" – apart from the transition structure and the spatiotemporal information – whose purpose would be to code which of the combinatorically possible histories are there in the structure. However, so far all attempts at providing such a coding have been inadequate. This points us to the first problem we want to put forward in the current paper:

Problem 1 *Create a format for discrete representation of arbitrary BST structures, such that a structure isomorphic to the original one may be recovered on*

9 See Thomas Müller, 2010, "Towards a Theory of Limited Indeterminism in Branching Space-Times", in: *Journal of Philosophical Logic* 39, pp. 395-423.

the basis of the representation. This would generalise Müller's 2010 theorem to arbitrary BST structures.

Again, in this short paper we cannot introduce Müller's approach.[10] Let us just note that we believe the above problem belongs to an interesting type: it looks really easy,[11] but has so far proven to be resistant to the attempts at solving it. Perhaps this points out that we still do not understand some basic facts about transition structures in BST.

To recapitulate, in the move from a BST structure to its transition structure there is a loss of information which prevents the move in the opposite direction: two non-isomorphic BST structures may have identical transition structures. Solving problem 1 requires a rigorous description of that loss. It is possible that using the tools from modal logic, and more specifically the notion of bisimulation, will be fruitful in that regard.[12]

Suppose a propositional language is given with a single modal operator. Each BST structure $\langle W, \leqslant\rangle$ may be regarded as a modal frame, with $\leqslant$ being the accessibility relation. A *BST model* is a triple $\langle W, \leqslant, V\rangle$, where $\langle W, \leqslant\rangle$ is a BST structure, and V is a valuation on W which in a sense does justice to the splitting inherent in the structure.[13] The following defines the notion of a bisimulation between two BST models:[14]

Definition 6 Suppose $\mathcal{M} = \langle W, \leqslant, V\rangle$ and $\mathcal{M}' = \langle W', \leqslant', V'\rangle$ are two BST models. A non-empty relation $Z \subseteq W \times W'$ is a *bisimulation* between $\mathcal{M}$ and $\mathcal{M}'$ when it satisfies the following three conditions, for any $w \in W$, $w' \in W'$:

- if wZw', then w and w' satisfy the same propositional letters;
- for any $v \in W$, if wZw' and $w \leqslant v$, then there exists a $v' \in W'$ such that vZv' and $w' \leqslant' v'$;
- for any $v' \in W'$, if wZw' and $w' \leqslant' v'$, then there exists a $v \in W$ such that vZv' and $w \leqslant v$.

Z is called a *total* bisimulation if for any $w \in W$ there exists a $w' \in W'$ such that wZw', and for any $w' \in W'$ there exists a $w \in W$ such that wZw'.

10 Building the formal machinery needed for this required many pages in *ibid.*

11 This opinion is based on conversations with other people accustomed with the BST framework.

12 We are very grateful to the Reviewer for suggesting this.

13 For details, see e.g. Tomasz Placek and Nuel Belnap, 2010, "Indeterminism is a Modal Notion: Branching Spacetimes and Earman's Pruning", in: *Synthese*, DOI 10.1007/s11229-010-9846-8.

14 For a description of the notion in the general context of modal logic, see e.g. Patrick Blackburn, Maarten de Rijke and Yde Venema, 2001, *Modal Logic*, Cambridge: Cambridge University Press.

Remember that in this paper all points in BST structures have space-time locations and all histories in all BST structures are isomorphic to the Minkowski space-time.

Conjecture 1 *Suppose* $OW = \langle W, \leqslant \rangle$ *and* $OW' = \langle W', \leqslant' \rangle$ *are two BST structures. Then* $TR(OW) = TR(OW')$ *iff there exist two BST models* $\mathcal{M} = \langle W, \leqslant, V \rangle$ *and* $\mathcal{M}' = \langle W', \leqslant', V' \rangle$ *with a total bisimulation* Z *between them such that for any* $w \in W$, $w' \in W'$, *if* wZw', *then* w *and* w' *have the same space-time location.*

Solving problem 1 would go a long way towards establishing the connection between BST and computer science. We aim to show this in the next two sections. Computer science contains numerous approaches to concurrent behaviour.[15] Of these we choose two: *event structures* and *Chu spaces*.

3. Event structures (for concurrent processes)

We will follow the presentation of the event structures framework from a paper by Varacca, Völzer and Winskel.[16]

Definition 7 An *event structure* (ES) is a triple $\mathcal{E} = \langle E, \leqslant, \# \rangle$ such that:

- E is countable;
- $\leqslant$ is a backward-finite partial order on E;
- $\#$ is an irreflexive and symmetric relation (the *conflict* relation) such that for every $e_1, e_2, e_3 \in E$, if $e_1 \leqslant e_2$ and $e_1 \# e_3$, then $e_2 \# e_3$.

The authors speak of E as the set of *events* and of $\leqslant$ as the *causal* order.

Notice the similarity of the conflict relation from ES with the modal branching of BST: if two events are in conflict, their descendants are also in conflict; if two BST histories branch, they never converge again.

If not for the requirement of backward-finitude, all BST structures $\langle OW, \leqslant \rangle$ and also their transition structures $\langle TR(OW), \leqslant_T \rangle$ would be ESs:

- for two BST events e, f, we put $e \# f$ iff $\neg \exists_{h \in Hist} e, f \in h$;
- for two transitions t_1, t_2 we put $t_1 \# t_2$ iff $\{t_1, t_2\}$ is inconsistent.

15 For an overview see Vaughan Pratt, 2003, "Transition and Cancellation in Concurrency and Branching Time", in: *Mathematical Structures in Computer Science* 13, 4, pp. 485-529.

16 Daniele Varacca, Hagen Völzer, and Glynn Winskel, 2006, "Probabilistic Event Structures and Domains", in: *Theoretical Computer Science* 358, 2-3, pp. 173-199.

The BST structures we are dealing with are all uncountable (even if they have just a single history, it is isomorphic to the Minkowski space-time), but their transition structures may well be countable and backward finite. In general, it is clear that any BST structure OW whose $TR(OW)$ is countable and backward finite determines an ES. This prompts a question about the other direction.

Problem 2 *What are the conditions for an event structure to be isomorphic to a transition structure $TR(OW)$ for some BST structure OW?*

Notice that definitely not all ESs are "suitable", because in BST we do not have "trivial" transitions, i.e. transitions whose first element would be an event which is not a choice point. And so consider an ES consisting of just a single event with the empty conflict relation: it lacks a natural BST reading, since there is no BST structure with just a single transition.[17]

3.1 Funny business causes confusion

It is interesting that the framework of event structures has a notion, called *confusion*, which seems to be similar to the BST notion of modal unsaturation (or "funny business").

In the following assume that a *configuration* of an ES $\mathcal{E}$ is a conflict-free downward closed subset of E. (And so maximal configurations in ESs correspond to the BST histories.) Also, define $[e] := \{x | x \leqslant e\}$ and $[e) := [e] \setminus \{e\}$.

Definition 8 Events e_1 and e_2 are in *immediate conflict* ($e_1 \#_\mu e_2$) when $e_1 \# e_2$ and both $[e_1) \cup [e_2]$ and $[e_1] \cup [e_2)$ are configurations.

A set of events C is a *partial cell* if for any distinct $e, e' \in C$ $e \#_\mu e'$ and $[e) = [e')$. A *cell* is a maximal partial cell.

We hope the reader will share the intuition that the ES notion of immediate conflict is similar in spirit to the BST notion of blatant inconsistency. It would also seem that all transitions from a single choice point to its immediate possibilities should, after the move from BST to ES, form a cell. This however may not be true if the BST structure exhibits modal unsaturation. The following table depicts the simple example of two BST structures having two binary choice points (and so the corresponding event structures are just four-element anti-chains), such that the first is exactly the modally saturated one depicted on p. 139, and the second one lacks one history. Notice that since the ESs are anti-chains, immediate conflict is just the "regular" conflict in this example.

In the first case the conflict relation joins only the elements corresponding to the transitions which are blatantly inconsistent in BST, and so the cells in the ES correspond to the sets of all transitions from a given choice point in the BST structure.

17 As a side-note, the following is a problem stated only once the connection between BST and ES has been noticed, but which relates to the current lack of deeper understanding of some fundamental aspects of BST: what are the sufficient and necessary conditions for a set of transitions to be the set of all transitions for some BST history?

BST:	**ES:**
Two SLR binary choice points e, f; 4 transitions	the anti-chain $\{e^+, e^-, f^+, f^-\}$
modal saturation	$e^+\#e^-$, $f^+\#f^-$ 2 cells: $\{e^+, e^-\}$,$\{f^+, f^-\}$.
modal unsaturation: the "++" history excluded	3 cells: $\{e^+, e^-\}$,$\{f^+, f^-\}$, and $\{e^+, f^+\}$.

However, when modal unsaturation enters the picture, the cells are no longer disjoint and closed under $\#_\mu$, thus losing their intuitive BST interpretation! It turns out that the ES framework has a notion pertaining to such cases:

Definition 9 An ES is *confusion free* if all its cells are closed under immediate conflict.

The above discussion prompts the following conjecture:

Conjecture 2 *Suppose OW is a BST structure such that $TR(OW)$ is countable and backward finite. Then OW is modally saturated iff $TR(OW)$ understood as an ES is confusion-free.*

If the conjecture is true, we will have a mathematical link between intuitively different concepts of "lack of a combinatorically possible option" (BST) and "weird choice structure" (ES).

To conclude this section: while it is easy how (if the cardinality requirements are met) to generate an ES isomorphic to the transition structure of a given BST structure, we are searching for a general method of proceeding in the other direction.

4. Chu spaces

The last framework to be considered is that of Chu spaces, for which our main reference is a paper by Pratt.[18] The Chu spaces are objects which are simple to define, but possess some great mathematical properties. To quote Pratt:[19] they form a "remarkably well-endowed category, concrete and coconcrete, self-dual, bicomplete, and symmetric monoidal closed", serving as "a process algebra representation of linear logic", "unifying relational structures, topology, and duality into a unified framework", providing a "process interpretation of wavefunctions" and (!) "*a solution to Descartes' problem of the mechanism by which the mind interacts with*

18 Vaughan Pratt, 1995, "Chu Spaces and Their Interpretation as Concurrent Objects", in: *Computer Science Today* 1000, pp. 392-405 (2005 version from the Author's homepage, http://boole.stanford.edu/pub/chuconc.pdf).

19 *Ibid.*, p. 3

the body" (emphasis added). Despite all this richness, for our goals it will suffice to think of Chu spaces as two-dimensional matrices.[20]

Definition 10 A *Chu space* over a set K is an $A \times X$ matrix whose elements are drawn from K.

In all the Chu spaces we will consider the set K is equal to $\{0, 1\}$.

Despite the apparent simplicity, the framework carries with itself robust interpretations of rows and columns of the matrices. If we view a given space by rows, then A is the "carrier of structure";[21] a row labeled e is the complete description of the element e. If we view it by columns, A is a set of "locations" (variables) and each column is a permitted assignment of values from K to them.[22] In our case rows will be labeled by transitions, and the "permitted assignments" will correspond to the characteristic functions of consistent sets of transitions (and the empty set).

Formally, the following is a method of representing BST structures by means of Chu spaces (notice the lack of cardinality requirements):

- there is a $1 - 1$ correspondence between the rows of the space and the transitions of the BST structures, which serve as labels for the rows;
- each column codes a possible past of an event in the BST structure, with "1" at all and only the rows whose corresponding transitions already happened from the perspective of the given point.

4.1 BST structures and "their" Chu spaces

The above will hopefully be made clearer by a few examples. We will always omit column labels. The Chu space corresponding to a BST structure with a single binary choice point (and so two transitions, labelled e^+ and e^-) is the following:

e^+	0	1	0
e^-	0	0	1

There are three columns because each event in the BST structure has one of the three possible pasts: it may be so that from its perspective e^+ already happened, or that e^- did, or none of those happened (yet). Since it is impossible for an event to have both e^+ and e^- in its past, as the two transitions are blatantly inconsistent, there is no column with two 1's.

A modally saturated BST structure with two binary choice points e and f (e.g. the one depicted on p. 139) gives rise to the following Chu space:

20 The mathematical value of Chu spaces seems to stem from the so called "Chu transforms", not introduced in this paper.

21 *Ibid.*

22 Perhaps the Reader will find the following quote illuminating: "The rows present the physical, concrete, conjunctive, or yang aspects of the space, while the columns present the mental, coconcrete, disjunctive, or yin aspects" (*ibid.*, p. 4).

e^+	0	1	0	0	0	1	1	0	0
e^-	0	0	1	0	0	0	0	1	1
f^+	0	0	0	1	0	1	0	1	0
f^-	0	0	0	0	1	0	1	0	1

The double line after the fifth column serves just to mark the point behind which each the columns determine which combinatorically possible histories are there in the model. Since the structure is modally saturated, all 4 possible histories are there.

The simplest case of modal unsaturation, with just a single history ("++") excluded, amounts just to the deletion of one column:

e^+	0	1	0	0	0	1	0	0
e^-	0	0	1	0	0	0	1	1
f^+	0	0	0	1	0	0	1	0
f^-	0	0	0	0	1	1	0	1

Notice that if we deleted all the columns in which two transitions happened, we would get the following:

e^+	0	1	0	0	0
e^-	0	0	1	0	0
f^+	0	0	0	1	0
f^-	0	0	0	0	1

which naturally represents a BST structure with just a single choice point with 4 immediate possibilities.

The last example should suggest that the relationship between BST structures and Chu spaces is not entirely straightforward. To reinforce this point, notice that not all Chu spaces over {0, 1} have a natural BST reading. Consider the following:

e^+	0	1	0	0	0	1
e^-	0	0	1	0	0	0
f^+	0	0	0	1	0	0
f^-	0	0	0	0	1	1

It cannot represent a structure with a single choice point, because the last column indicates that it is possible to have two transitions in the past. But there seems to be no way of looking at this space as representing a BST structure with two or any other number of choice points.

Notice that the above examples show that sometimes, starting from a Chu space representing a modally saturated BST structure, one can, just by deleting columns which seemingly correspond to combinatorically possible histories, introduce modal unsaturation, then lose the natural BST reading altogether, and eventually end up with a space representing a BST structure with a different number of choice points. Contrast this with the process of removing histories from the modally saturated BST structure depicted on p. 139: if we remove one history (say

h_2), we introduce modal unsaturation, if we remove two histories (say h_2 and h_3; notice that not all choices are permissible), we again have a structure with modal unsaturation, if we remove three histories we get a structure with no choice points at all, and if we remove all four histories we end up with the empty set. Perhaps a general theorem on the relationship between BST structures and Chu spaces requires a deeper understanding of the connection between columns representing the "latest possible pasts" (i.e. the right-most columns in our examples) and the combinatorically possible histories in BST structures.

We can, however, provide a procedure which given a Chu space $A' \times X'$ over $\{0, 1\}$ (with A' countable) creates an event structure $\langle E, \leqslant, \# \rangle$ (whenever $\leqslant$ turns out to be backward-finite):

1. Delete any repeated rows from A' and columns from X' (save for a single copy in each case), arriving at A and X.

2. Set E to be A.

3. For the ordering $\leqslant$, take the bit-wise ordering of rows, given by the inverse of the "left residual" of $A \times X$ and itself: namely, the set of pairs $\langle b, a \rangle$ of elements of A such that for any column $x \in X$, if there's a 1 at row a and column x, then there is a 1 at row b and column x (in such a case we want to say that $b \leqslant a$).

4. Set $e \# f$ for any and all $e, f \in A$ such that no column contains 1's at both rows e and f.

Were we able to prove the theorem about discrete representations of BST structures in full generality (see Problem 1), we could move all the way from Chu spaces, via event structures, to BST structures. As things stand, the known method[23] of constructing a BST structure on the basis of a given transition structure always generates a modally saturated BST structure.

We are left with a similar problem as in the case of event structures:

Problem 3 *What are the conditions for a Chu space over* $\{0, 1\}$ *to generate a transition structure* $TR(OW)$ *for some BST structure* OW*?*

5. Conclusion

In this paper we put forward two conjectures and three problems regarding the relationship of BST structures, event structures for concurrent processing and Chu spaces. It seems to be relatively easy to generate the latter objects given BST structures (preserving what we believe to be important: the shape of the transition structure), and more difficult to move in the other direction.

23 See Müller, "Towards a Theory of Limited Indeterminism in Branching Space-Times", *op. cit*.

We hope that in the process of investigating these problems we will gain some insight into the relationship between concepts from seemingly unrelated fields of philosophy and computer science, between which nonetheless there definitely seems to be a mathematical connection: as an example, take the notion of modal unsaturation (BST) and confusion (ES), the topic of Section 3.1.

The investigation so far suggests that there is much to gain in this for BST theorists – using the tools from computer science may offer a new look at some BST problems and provide a better understanding of transition structures. Still, perhaps some computer scientists will also be interested in spatiotemporal readings of their structures.

Acknowledgements: We would like to thank the Reviewer for fruitful comments and Thomas Müller for suggesting the topic in the first place and sharing his thoughts on the key issues of the paper. Our thanks also go to Vincent van Oostrom for a brainstorming discussion. The research was supported by the MNiSW grant no. 668/N-RNP-ESF/2010/0.

Institute of Philosophy
Jagiellonian University
Grodzka 52
31-044, Kraków
Poland
leszek.wronski@uj.edu.pl

Team B
Philosophy of Systems Biology

Gabriele Gramelsberger

Simulation and System Understanding

Abstract

Systems biology is based on a mathematized understanding of molecular biological processes. Because genetic networks are so complex, a system understanding is required that allows for the appropriate modelling of these complex networks and its products up to the whole-cell scale. Since 2000 standardizations in modelling and simulation techniques have been established to support the community-wide endeavors for whole-cell simulations. The development of the Systems Biology Markup Language (SBML), in particular, has helped systems biologists achieve their goal. This paper explores the current developments of modelling and simulation in systems biology. It discusses the question as to whether an appropriate system understanding has been developed yet, or whether advanced software machineries of whole-cell simulations can compensate for the lack of system understanding.

1. Towards a simulation-orientated biology

In a 2002 *Nature* paper systems biology was defined as the "mathematical concepts […] to illuminate the principles underlying biology at a genetic, molecular, cellular and even organismal level."[1] During the past years these mathematical concepts have become 'whole-cell simulations' in order to observe, and hopefully understand, the complex dynamic behavior of cells. Already in 1997 the very first minimal cell was created in-silico, called the 'virtual self-surviving cell (SSC)', consisting of 120 in-silico synthesized genes of the 480 genes of *M. genitalium* and 7 genes from other species.[2] The virtual self-surviving cell absorbs up and metabolizes glucose, and generates ATP as an energy source for protein and membrane synthesis. As degradation is programmed into the SSC, it has to produce proteins and lipids constantly to survive. All the activities result from 495 reaction rules for enzymatic reactions, complex formations, transportations, and stochastic processes, which are executed in parallel telling the SSC what to do at each millisecond time step. The aims of this whole-cell simulation are to observe the changes in the

1 Christopher Surridge (Ed.), "Nature inside: Computational Biology", in: *Nature* 420, 2002, 205-250, here p. 205.

2 Cf. Masaru Tomita, "Whole-cell Simulation: A Grand Challenge of the 21st Century", in: *TRENDS in Biotechnology* 19, 6, 2001, pp. 205-210.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_13,

amount of substances inside the cell as well as the gene expression resulting from these changes, to study the temporal patterns of change, and, finally, to conduct experiments with the SSC, e.g. real-time gene knock-out experiments. Thus,

> [t]his simple cell model sometimes shows unpredictable behavior and has delivered biologically interesting surprises. When the extracellular glucose is drained and set to be zero, intracellular ATP momentarily increases and then decreases […]. At first, this finding was confusing. Because ATP is synthesized only by the glycolysis pathway, it was assumed that ATP would decrease when the glucose, the only source of energy, becomes zero. After months of checking the simulation program and the cell model for errors, the conclusion is that this observation is correct and a rapid deprivation of glucose supplement can lead to the same phenomenon in living cells.[3]

The motivation of making use of simulation in biology is the desire to predict effects of changes in cell behavior and guide further research in the lab. As metabolic networks are extremely complex systems involving dozens and hundreds of genes, enzymes, and other species, it is difficult to study these networks experimentally. Therefore, in-silico studies are increasingly expanding the wet lab studies, but this requires the full range of strategies necessary to establish a simulation-orientated biology. These strategies are standardization, acquisition of sufficient spatiotemporal information about processes and parameters, creation of advanced software machineries, and, last but not least, a coherent system understanding.

2. Standardization

The situation of modeling and simulation in cell biology is characterized by a wide variety of modeling practices and methods. There are thousands of simple models around, an increasing amount of simulations for more complex models, and various computational tools to ease modeling. Each institute, each researcher creates his or her own model with slightly different concepts and meanings. Most of these models are not comparable with each other, because "each author may use a different modeling environment (and model representation language), [therefore] such model definitions are often not straightforward to examine, test and reuse."[4] However, in 2000 this situation led to an effort to create an open and standardized framework for modeling – the Systems Biology Markup Language (SBML). The collaborative work on SBML was motivated by the goal to overcome "the current inability to exchange models between different simulation and analysis tools [which] has its roots in the lack of a common format for describing models."[5]

3 Tomita 2001, *loc. cit*., here p. 208.

4 Michael Hucka et al., "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models", in: *Bioinformatics* 19, 2003, pp. 524–531, here p. 525.

5 Hucka et al. 2003, *loc. cit*., here p. 524. SBML is organized as a community-wide open

Therefore, SBML is based on the Extensible Markup Language (XML), a set of rules for encoding documents in machine-readable form, developed in 1996 by the World Wide Web Consortium. After three years of work with many contributions from the community, in particular from teams developing simulation and analysis packages for systems biology, SBML level 1 was released in 2003. As is the nature of software development, level 2 was tackled immediately after the release of level 1 containing more features; since 2010, with level 3, SBML has practically become a lingua franca of model description in biology – according to *Nature*'s website: "There are now hundreds of software systems supporting SBML; further, many journals now suggest it as a format for submitting supplementary data."[6]

The advantage of SBML is the standardization of modeling by defining conceptual elements of chemical reactions. These elements are compartments, species, reactions, parameters, unit definitions, and rules. For instance, a hypothetical single-gene oscillatory circuit can be modeled with SBML as a simple, two compartment model: one compartment for the nucleus, one for the surrounding cell cytoplasm.[7] This circuit can produce nine species and each species is defined by its 'name', by its 'initialAmount', and optionally by 'boundaryCondition' and 'charge'. The species produced result from the eight reactions which the hypothetical single-gene circuit is capable of. "In SBML, reactions are defined using lists of reactant species and products, their stoichiometric coefficients, and kinetic rate laws."[8] This means that every reaction has to be specified by its 'name', by the 'species' involved as reactants or products, by the attribute 'reversible', which can have the values false or true, and by the optional attribute 'fast'. If the attribute 'fast' has the value true, "simulation/analysis packages may choose to use this information to reduce the number of [ordinary differential equations] ODEs required and thereby optimize such computations."[9] Finally, 'rules' can be set to constrain variables and parameters. However, SBML is solely a format to describe a model. For simulation it has to be transferred to simulation and analysis packages that support SBML. In 2003 nine simulation tools have supported SBML (*Cellerator*, *DBsolve*, *E-CELL*, *Gepasi*, *Jarnac*, *NetBuilder*, *ProMoT/DIVA*, *StochSim*, and *Virtual Cell*), while today more than two hundred packages do. Based on the success of SBML, Systems Biology Graphical Notation (SBGN) has recently been developed – and released in 2009 – as a community-wide open graphical standard that allows three different views of biological systems: process descriptions, en-

standard based on open workshops and an editorial team for updates, which is elected to a 3-year non-renewable term. (Cf. http://sbml.org/).

6 Nature Precedings: *The Systems Biology Markup Language (SBML) Collection*, (accessed on 6 January 2012). URL: http://precedings.nature.com/collections/sbml. Another description language for modeling is CellML.

7 The example, and its notations, is taken from the initial SBML paper. Cf. Hucka et al., 2003, *loc. cit*., p. 526 ff.

8 *Ibid*., p. 528.

9 *Ibid*., p. 528.

tity relationships and activity flows, creating three different types of diagrams. Furthermore, new software applications, like *CellDesigner* and *CellPublisher*, are now built on both SBML and SBGN.

3. Quantitative data

Another important basis of preparing the stage for a simulation-orientated biology is the acquisition of quantitative data as most simulation methods in biology are based on differential equations, which describe the temporal development of a system's behavior. However, most data available in biology are qualitative data representing the functions of genes, pathway maps, protein interaction, etc. "But for simulation quantitative data (such as concentrations of metabolites and enzymes, flux rates, kinetic parameters and dissociation constants) are needed. A major challenge is to develop high-throughput technologies for measurement of inner-cellular metabolites."[10] Furthermore, these quantitative data have to be fine-grained. "[T]raditional biological experiments tend to measure only the change before and after a certain event. For computational analysis, data measured at a constant time interval are essential in addition to traditional sampling points."[11] However, quantitative measurements of inner-cellular metabolites, protein synthesis, gene expression, etc. in fine-grained time series experiments are challenging experimental biology. For instance, it is assumed that eukaryotic organisms contain between 4,000 and 20,000 metabolites. Unlike proteins or RNA, the physical and chemical properties of these metabolites are highly divergent and there is a "high proportion of unknown analytes that is measured in metabolite profiling. Typically, in current [Gas-chromatography–mass-spectrometry] GC–MS-based metabolite profiling, a chemical structure has been unambiguously assigned in only 20–30% of the analytes detected."[12] Nevertheless, a typical GC-MS profile of one sample contains 300 to 500 analytes and generates a 20-megabyte data file. Quantitative data for metabolite profiles as well as for transcript and protein profiling are usually expressed as ratios of a control sample. "In addition, absolute quantification is important for understanding metabolic networks, as it is necessary for the calculation of atomic balances or for using kinetic properties of enzymes to develop predictive models."[13] For both types of quantitative data the changes in the samples, even for tiny influences, can be huge and it is difficult to achieve meaningful and reliable

10 Tomita 2001, *loc. cit*., here p. 210. Cf. Jörg Stelling, et al., "Towards a Virtual Biological Laboratory", in: Hiroaki Kitano (Ed.), *Foundations of Systems Biology*. Cambridge (Mass.): The MIT Press 2001, pp. 189-212.

11 Hiroaki Kitano, "Systems Biology: Toward System-level Understanding of Biological Systems", in: Hiroaki Kitano, 2001, *op. cit*., pp. 1-38, here p. 6.

12 Alisdair R. Fernie, et al., "Metabolite Profiling: From Diagnostics to Systems Biology", in: *Nature Reviews Molecular Cell Biology* 5, 2004, pp. 763-769, here p. 764.

13 Fernie et al., 2004, here p. 765.

results. Furthermore, these results do not come from fine-grained time series, let alone cross-category measurements of "metabolites, proteins and/or mRNA from the same sample […] to assess connectivity across different molecular entities."[14]

However, fine-grained quantitative information is required for simulation in biology. The answers to this challenge are manifold. One approach is to address the measurement problem by creating new facilities, methods, and institutes. For instance, in Japan a new institute has been set up "for this new type of simulation-orientated biology, [… which] consists of three centers for metabolome research, bioinformatics, and genome engineering, respectively."[15] Or alternatively, other simulation methods for specific purposes have to be used "to deal with the lack of kinetic information, […] for instance flux balance analysis (FBA)." [16] FBA does not require dynamic data as it analyzes the capabilities of a reconstructed metabolic network on basis of systemic mass-balance and reaction capacity constraints. Yet flux balance analysis does not give a unique solution for the flux distribution – only an optimal distribution can be inferred.[17] Another alternative could be the estimation of unmeasurable parameters and variables by models. So-called non-linear state-space models have been developed recently for the indirect determination of unknown parameters from measurements of other quantities. These models take advantage of knowledge that is hidden in the system, by training the models to learn more about themselves.

4. Whole-cell simulations

Whatever option is chosen to tackle the problems that come along with a simulation-orientated biology, whole-cell simulations show great promise. On the one hand they are needed for data integration,[18] on the other hand the "ultimate goal

14 *Ibid*., p. 768.

15 Tomita 2001, *loc. cit*., here p. 210.

16 J. S. Edwards, R. U. Ibarra, B. O. Palsson, "In Silico Predictions of Escherichia coli Metabolic Capabilities are Consistent with Experimental Data", in: *Nature Biotechnology* 19, 2001, pp. 125–130, here p. 125.

17 "As a result of the incomplete set of constraints on the metabolic network (that is, kinetic constant constraints and gene expression constraints are not considered), FBA does not yield a unique solution for the flux distribution. Rather, FBA provides a solution space that contains all the possible steady-state flux distributions that satisfy the applied constraints. Subject to the imposed constraints, optimal metabolic flux distributions can be determined from the set of all allowable flux distributions using linear programming (LP)." (Edwards, Ibarra, Palsson, 2001, *loc cit*., here p. 125).

18 "[…] a crucial and obvious challenge is to determine how these, often disparate and complex, details can explain the cellular process under investigation. The ideal way to meet this challenge is to integrate and organize the data into a predictive model." (Boris M. Slepchenko et al., "Quantitative Cell Biology with the Virtual Cell", in: *TRENDS in Cell Biology* 13, 11, 2003, pp. 570-576, here p. 570). Olaf Wolkenhauer

[…] is to construct a whole-cell model in silico […] and then to design a novel genome based on the computer simulation and create real cells with the novel genome by means of genome engineering;"[19] in brief: to enable the "computer aided design (CAD) of useful microorganisms."[20] Projects like *E-Cell* or *Virtual Cell* – just to mention two – aim "to develop the theories, techniques, and software platforms necessary for whole-cell-scale modeling, simulation, and analysis."[21] *E-Cell* and *Virtual Cell* differ in their conception as well as organization. Development of the *E-Cell* software in C++ started in 1996 at the Laboratory for Bioinformatics at Keio University, initiated by Masaru Tomita. In 1997 the 1.0beta version was used to program the 'virtual self-surviving cell', which was accepted as an OpenSource project by the Bioinformatics.org in 2000. The software development led to the establishment of the Institute for Advanced Biosciences for metabolome research, bioinformatics, and genome engineering in 2001 and by 2005 the Molecular Sciences Institute, Berkeley, and the Mitsubishi Space Co. Ltd, Amagasaki Japan, had joined the project. *E-Cell* is used internationally by various research groups to realize in-silico projects, e.g. on the dynamics of mitochondrial metabolism, on the energy metabolism of *E. coli*, on glycolysis, etc.[22] The software as well as the already programmed models can be downloaded from the web page. Unlike *E-Cell*, *Virtual Cell* is a freely accessible software platform by the Center for Cell Analysis & Modeling of the University of Connecticut for building complex models with a web-based Java interface. Thus, the "mathematic-savy user may directly specify the complete mathematical description of the model, bypassing the schematic interface."[23] *Virtual Cell* is conceived as an open community platform, providing software releases and programmed models. Thus, the increasing availability of an open community cyberinfrastructure for a simulation-orientated biology can be observed as is already common for other disciplines like meteorology.[24]

---

and Ursula Klingmüller have expanded the definition of systems biology given in the first paragraph of this paper by adding "the integration of data, obtained from experiments at various levels and associated with the 'omics family' of technologies." (Olaf Wolkenhauer, Ursula Klingmüller, "Systems Biology: From a Buzzword to a Life Science Approach", in: *BIOforum Europe* 4, 2004, pp. 22-23, here p. 22).

19 Tomita 2001, *loc. cit*., here p. 210.

20 Masaru Tomita, "Towards Computer Aided Design (CAD) of Useful Microorganisms", in: *Bioinformatics* 17, 12, 2001a, pp. 1091-1092.

21 Kouichi Takahashi et al., "Computational Challenges in Cell Simulation: A Software Engineering Approach", in: *IEEE Intelligent Systems* 5, 2002, pp. 64-71, here p. 64.

22 Cf. E-Cell: Homepage, (accessed on 6 January 2012). URL: http://www.e-cell.org/ecell.

23 Virtual Cell: Homepage at the Center for Cell Analysis & Modeling, (accessed on 6 January 2012). URL: http://www.nrcam.uchc.edu/.

24 Cf. Gabriele Gramelsberger, Johann Feichter, "Modeling the Climate System", in: Gabriele Gramelsberger, Johann Feichter (Eds.), *Climate Change and Policy. The Calculability of Climate Change and the Challenge of Uncertainty*. Heidelberg: Springer 2011, p. 44 ff.

However, the aim of this simulation approach is the creation and distribution of complex models. These collaboratively advanced software machineries are 'synthesizers' for interconnecting all kinds of computational schemes and strategies. Thus, complex systems are built in a bottom-up process by innumerous computational schemes. *E-Cell*, for example, combines object-oriented modeling for DNA replication, Boolean networks and stochastic algorithms for gene expression, differential-algebraic equations (DAEs)[25] and FBA for metabolic pathways, SDEs and ODEs for other cellular processes (see Tab. 1).[26] These advanced integrative cell simulations provide in-silico experimental devices for hypothesis testing and predictions, but also for data integration and the engineering of de-novo cells. In such an in-silico experimental device genes can be turned on and off, substance concentrations and metabolites can be changed, etc. Observing the behavior of the in-silico cell yields comparative insights and can lead to the discovery of causalities and interdependencies. Thus, "computers have proven to be invaluable in analyzing these systems, and many biologists are turning to the keyboard."[27] Researchers involved in the development of the whole-cell simulator *E-Cell* have outlined a 'computational cell biology research cycle' from wet experiments forming cellular data and hypotheses, to qualitative and quantitative modeling, to programming and simulation runs, to analysis and interpretation of the results, and back to wet experiments for evaluation purposes.[28] However, this is still a vision as "biologists rarely have sufficient training in the mathematics and physics required to build quantitative models, [therefore] modeling has been largely the purview of theoreticians who have the appropriate training but little experience in the laboratory. This disconnection to the laboratory has limited the impact of mathematical modeling in cell biology and, in some quarters, has even given modeling a poor reputation."[29] In particular *Virtual Cell* tries to overcome this by offering an intuitive modeling workspace that is "abstracting and automating the mathematical and physical operations involved in constructing models and generating simulations from them."[30]

25 A DAE combines one ordinary differential equation (ODE) for each enzyme reaction, a stochimetric matrix, and algebraic equations for constraining the system.

26 Cf. Takahashi et al., 2002, *loc. cit*., p. 66 ff.

27 *Ibid*., p. 64.

28 Cf. *Ibid*., p. 64 ff.

29 Leslie M. Loew, et al., "The Virtual Cell Project", in: *Systems Biomedicine*, 2010, pp. 273-288, here p. 274.

30 Loew, et al., 2010, *loc. cit*., here p. 274.

Tab.1 **Cellular processes and typical computational approaches** (replotted from Takahashi et al., 2002, p. 66)

Process type	Dominant phenomena	Typical computation schemes
Metabolism	Enzymatic reaction	DAE, S-Systems, FBA
Signal transduction	Molecular binding	DAE, stochastic algorithms (e.g. StochSim , Gillespie), diffusion-reaction
Gene expression	Molecular binding, polymerization, degradation	OOM, S-Systems, DAE, Boolean networks, stochastic algorithms
DNA replication	Molecular binding, polymerization	OOM, DAE
Cytoskeletal	Polymerization, depolymerization	DAE, particle dynamics
Cytoplasmic streaming	Streaming	Rheology, finite-element method
Membrane transport	Osmotic pressure, membrane potential	DAE, electrophysiology

5. System understanding

However, the core of systems biology is a proper system understanding, which can be articulated mathematically and modeled algorithmically. As the systems biologist Hiroaki Kitano pointed out in 2001, "systems biology aims at understanding biological systems at system level,"[31] referring to forerunners like Norbert Wiener and Ludwig von Bertalanffy. For Bertalanffy "a system can be defined as a complex of interacting elements."[32] However, as biology uses mathematical tools and concepts developed in physics, it has to be asked whether the applied computational schemes suit biological systems. The problem with the concept of 'complex systems' resulting from physics is twofold. On the one hand physical complex systems are characterized by elements "which are the same within and outside the complex; they may therefore be obtained by means of summation of characteristics and behavior of elements as known in isolation" (summative characteristic).[33] On the other hand their concept of interaction exhibits an unorganized complexity, while the main characteristic of biological systems is their organized complexity.

31 Kitano 2001, *op cit*., here p. xiii referring to Norbert Wiener: *Cybernetics or Control and Communication in the Animal and the Machine*. New York: John Wiley & Sons 1948 and Ludwig von Bertalanffy, *General System Theory. Foundations, Development, Applications*. New York: Braziller 1968.

32 von Bertalanffy, 1968, *op. cit*., here p. 55.

33 Hiroaki Kitano, "Computational Systems Biology", in: *Nature* 420, 2002, pp. 206-210, here p. 54.

Therefore, Kitano calls biological systems 'symbiotic systems' exhibiting coherent rather than complex behavior:

> It is often said that biological systems, such as cells, are 'complex systems'. A popular notion of complex systems is of very large numbers of simple and identical elements interacting to produce 'complex' behaviours. The reality of biological systems is somewhat different. Here large numbers of functionally diverse, and frequently multifunctional, sets of elements interact selectively and nonlinearly to produce coherent rather than complex behaviours.
>
> Unlike complex systems of simple elements, in which functions emerge from the properties of the networks they form rather than from any specific element, functions in biological systems rely on a combination of the network and the specific elements involved. […] In this way, biological systems might be better characterized as symbiotic systems.[34]

In contrast to physics, biological elements are innumerous, functionally diverse, and interact selectively in coupled and feedbacked sub-networks, thus characterizing the properties of a biological system. The challenge for systems biology is: how to conceive biological complexity? In retrospect, the basic question of system understanding was already discussed in the 19th century's mechanism-vitalism debate.[35] In the 20th century the idea of self-regulating systems emerged and two different concepts have been developed against the complex system approach of physics: the steady-state (Fliessgleichgewicht = flux equilibrium) or open system concept by Bertalanffy and the feedback regulation concept of Wiener referring to Walter B. Cannon's biological concept of homeostasis.[36] Bertalanffy aptly describes the differences between both concepts. For him Cannon's homeostatic control and Wiener's feedback systems are special classes of self-regulating systems. Both are

> 'open' with respect to incoming information, but 'closed' with respect to matter. The concepts of information theory–particularly in the equivalence of information and negative entropy–correspond therefore to 'closed' thermodynamics (thermostatics) rather than irreversible thermodynamics of open systems. However, the latter is presupposed if the system (like the living organism) is to be 'self-organizing'. […]
>
> Thus dynamics in open systems and feedback mechanisms are two different model concepts, each right in its proper sphere. The open-system model is basically nonmechanistic, and transcends not only conventional thermodynamics, but also one-way causality as is basic in conventional physical theory. The cybernetic approach retains the Cartesian machine

34 Kitano, 2002, *loc cit*., here p. 206.

35 Cf. Ulrich Krohs, Georg Toepfer (Eds.), *Philosophie der Biologie*. Frankfurt: Suhrkamp 2005.

36 Cf. Ludwig von Bertalanffy, *Theoretische Biologie*. Berlin: Bornträger 1932; Ludwig von Bertalanffy, *Biophysik des Fließgleichgewichts*. Berlin: Akademie Verlag 1952; Wiener, 1948, *op. cit*.; Walter B. Cannon, "Organization for Physiological Homeostasis", in: *Physiological Review* 9, 1929, p. 397; Walter B. Cannon, *The Wisdom of the Body*. New York: Norton 1932.

model of the organism, unidirectional causality and closed systems; its novelty lies in the introduction of concepts transcending conventional physics, especially those of information theory. Ultimately, the pair is a modern expression of the ancient antithesis of 'process' and 'structure'; it will eventually have to be solved dialectically in some new synthesis.[37]

Following Bertalanffy's ideas, "the living cell and organism is not a static pattern or machine like structure consisting of more or less permanent 'building materials' […, but] an 'open system'."[38] While in physics systems are usually conceived as closed ones, sometimes expanded by additional terms for energy exchange with their environment, biological systems are open in regard to energy and matter. Therefore biological systems show characteristic principles of behavior like steady state, equifinality, and hierarchical organization. "In the steady state, the composition of the system remains constant in spite of continuous exchange of components. Steady states or *Fliessgleichgewicht* are equifinal […]; i.e., the same time-independent state may be reached from different initial conditions and in different ways – much in contrast to conventional physical systems where the equilibrium state is determined by initial conditions."[39] This can lead to 'overshoots' and 'false starts' as a response to unstable states and external stimuli (adaptation), because biological systems tend towards a steady state. Based on this concept the overshoot of intercellular ATP as exhibited by the virtual self-surviving cell (see Sect. 1) can be easily explained, while for a physicist it sounds mysteriously.

However, the basic question is: Does simulation support this new system understanding beyond the physical concept of complex systems? Can biological systems modeled based on non-summative characteristics, meaning that a complex is not built up "step by step, by putting together the first separate elements; […but by using] constitutive characteristics […] which are dependent on the specific relations within the complex; for understanding such characteristics we therefore must know not only the parts, but also their relations."[40] The brief overview of modeling practices with SBML has shown that each part (species, reaction, etc.) has to be described explicitly. However, the important aspect is the interaction between these parts (flux). Advanced software machineries allow complex interactions and feedbacks to be organized, e.g. feedbacks, loops, alternatives (jumps), etc. Thus, the software machineries of whole-cell simulations function as 'synthesizers' and can be seen as media for organizing complex relations and fluxes. As already defined by Herman Goldstine and John von Neumann in 1946, "coding begins with the drawing of the flow diagrams."[41] These flow diagrams specify the

37 von Bertalanffy 1986, *op. cit*., here p. 163.

38 *Ibid*., p. 158.

39 *Ibid*., p. 159.

40 *Ibid*., pp. 67 and 55.

41 Herman H. Goldstine, John von Neumann, "Planning and Coding Problems for an Electronic Computing Instrument" (1947), Part II, vol. 1, in: John von Neumann, *Collected Work*, vol. V: Design of Computers, Theory of Automata and Numerical Analy-

various sequences of calculations as well as routines that have to be repeated. The result is a complex choreography of command courses and loops; and "the relation of the coded instructions to the mathematically conceived procedures of (numerical) solutions is not a statical one, that of a translation, but highly dynamical."[42] The dynamics of the actual path of computing through the instructions (simulation run), which Goldstine and Neumann called the 'modus procedendi', requires an ambitious choreography of possibilities and alternatives expressed by 'if, then, else, end if'-decisions, loops, and calls of other subroutines within each singular file of a program. Moreover, object-oriented programming allows for 'simulating' complex, organizational structures by defining objects and classes with specific properties for varying contexts, imitating selectivity and functional diversity.[43] While the underlying mathematical tools and computational schemes, resulting from physics, haven't changed, advanced software machineries allow their 'organized complexity'. Thus, the simulation approach not just supports, but enables the new system understanding for biology. One can ask, if it only bypasses the core challenge of biological complexity by mimicking organized complexity. However, if this is the case, biology has to inspire a new mathematics, just as physics did four centuries ago by developing the calculus for describing the kinetics of spatio-temporal systems.

FU Institute of Philosophy
Freie Universität Berlin
Habelschwerdter Allee 30
D-14195, Berlin
Germany
gab@zedat.fu-berlin.de

sis, Oxford: Pergamon Press 1963, pp. 80-151, here p. 100. Cf. Gabriele Gramelsberger, "From Computation with Experiments to Experiments with Computation", in: Gabriele Gramelsberger (Ed.), *From Science to Computational Sciences. Studies in the History of Computing and its Influence on Today's Sciences*. Zurich: Diaphanes, pp. 131-142.

42 Goldstine, Neumann, 1947, *loc. cit*., here pp. 81-82.

43 Interestingly, the object-oriented programming paradigm – introduced in 1967 with Simula 67 for physical simulations – was advanced for the programming language C++, which originated in the telecommunication industry (Bell labs) in response to the demand for more complex structures for organizing network traffic. Cf. Terry Shinn, "When is Simulation a Research-Technology? Practices, Markets and Lingua Franca", in: Johannes Lenhard, Günter Küppers, Terry Shinn (Eds.), *Simulation: Pragmatic Construction of Reality*. Dordrecht: Springer, 2006, pp. 187-203.

Tarja Knuuttila and Andrea Loettgers

Synthetic Biology as an Engineering Science? Analogical Reasoning, Synthetic Modeling, and Integration

Abstract

Synthetic biology has typically been understood as a kind of engineering science in which engineering principles are applied to biology. The engineering orientation of synthetic biology has also received a fair deal of criticism. This paper presents an alternative reading of synthetic biology focusing on the basic science oriented branch of synthetic biology. We discuss the practice of synthetic modeling and how it has made synthetic biologists more aware of some fundamental differences between the functioning of engineered artifacts and biological organisms. As the recent work on the concepts of noise and modularity shows, synthetic biology is in the process of becoming more "biology inspired".

1. Introduction

Systems biology and synthetic biology form related, highly interdisciplinary fields sharing largely the same analytic tools. What sets them apart is the focus of synthetic biology on the design and construction of novel biological functions and systems. Synthetic biology is often understood in terms of the pursuit for well–characterized biological parts to create synthetic wholes,[1] and as such has typically been understood as a kind of engineering science in which engineering principles are applied to biology. This view is shared by the public understanding of synthetic biology as well as the practitioners themselves. According to Jim Collins[2], who introduced one of the first synthetic networks, a toggle-switch, in 2000: "[…] synthetic biology was born with the broad goal of engineering or 'wiring' biological circuitry – be it genetic, protein, viral, pathway or genomic – for manifesting logical forms of cellular control."

The engineering orientation of synthetic biology has received a fair deal of criticism. In a recently published article on systems and synthetic biology Calvert

1 Church, G. M., "From Systems Biology to Synthetic Biology", in: *Molecular Systems Biology* 1, 2005.0032, doi:10.1038/msb4100007, Published online: 29 March 2005.

2 Khalil, S. A. and Collins, J. J., "Synthetic Biology: Applications Come to Age", in: *Nature Reviews Genetics* 11, 2010, pp. 367–379.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_14,

and Fujimura[3] claim that "[t]he research programme that expresses this objective [of rendering life calculable] in perhaps its most extreme form is *synthetic biology*". Furthermore, they posit that "synthetic biology aims at construction, whereas the objective of systems biology is to understand existing biological systems" (*ibid*.). We wish to present an alternative reading of synthetic biology that pays attention to the epistemic dimension of the material practice of the discipline. Taking into account the impressive array of interview and other data on which Calvert and Fujimura's study was based, we find it astonishing that they neither recognize the basic science oriented approach of synthetic biology nor distinguish between the influences of engineering vis-à-vis physics on synthetic biology. Namely, a more basic science oriented branch of synthetic biology has developed alongside the more engineering and application oriented approaches. This basic science oriented branch of synthetic biology targets our understanding of biological organization by probing the basic "design principles" of life. The design and exploration of gene regulatory networks constructed from biological material and implemented in natural cell environment is exemplary of this kind of approach. Interestingly, this kind of study has directly affected synthetic biology: biology in all its complexity has begun to occupy the centre stage. Important engineering notions on which synthetic biology has been grounded, such as noise and modularity, have been reinterpreted and some analogies drawn to engineering have been questioned. In the following we will study some aspects of this development through consideration of work at the Elowitz lab, which is one of the leading synthetic biology laboratories.[4]

2. Analogical reasoning and combinatorial modeling

2.1 Physicists advertising the use of engineering concepts in biology

In synthetic biology one can distinguish two main approaches: an engineering approach and a basic science approach. The engineering approach, which aims to design novel biological parts or organisms for the production of, for instance,

3 Calvert, J. and Fujimura, J., "Calculating Life? Duelling Discourses in Interdisciplinary Systems Biology", in: *Studies in History and Philosophy of Biological and Biomedical Sciences,* 42, 2011, p. 160.

4 One of the authors spent four years in the Elowitz lab at the California Institute of Technology observing the daily research practice in this lab.

vaccines,[5] biofuels,[6] and cancer-killing bacteria,[7] is often construed as comprising the whole field of synthetic biology. Less visible than the engineering approach is the basic science approach, which uses synthetic biology, especially synthetically designed biological parts, as a tool for investigating the basic design principles of gene-regulatory networks.[8] When this line of research took its first steps, one of the main desiderata was to reduce the complexity of biological systems. The reason for this strategy was not necessarily due to the reductive vision of the scientists in question but rather their aim of studying some aspects of biological organization *in isolation*. This was deemed indispensable for the purposes of testing various possible design principles, as well as exploring the concepts, methods and techniques imported to systems and synthetic biology from other disciplines, notably from engineering and physics.

It is remarkable, in the first place, that engineers and physicists did start to experiment, explore, and tinker with biological systems. To be sure, there are plenty of examples throughout history, of physics and physicists having an important impact on theoretical work in biology. Yet, during the emergence of synthetic biology something rather new happened: physicists entered biology labs or even opened their own labs and started working at the bench. This movement of physicists into molecular biology labs was largely enabled by the standardized molecular biology kits, which became available by that time. With these kits, no longer was it essential to know all the details and steps of polymerase chain reactions (PCR) – a method to amplify a small number of copies of DNA – one could simply follow the instructions that came with the kit. Performing experiments in molecular biology was suddenly much easier. Another peculiar feature of synthetic biology is that even though the basic science approach has been heavily physics-influenced, many of the central concepts come from engineering. This raises the question of what triggered this use of engineering concepts by physicists. Why does one not immediately recognize "the physicist" behind this line of research?

Interestingly, physicists themselves have argued against the use of concepts taken from physics in describing and analyzing biological systems. Physicists

5 Ro, D. K., Paradise, E., Quellet, M., Fisher, K., Newman. K., Ndgundu, J., Ho, K., Eachus, R., Ham, T., Kirby, J., Chang M. C. Y., Withers, S., Shiba, Y., Sarpong, R. and Keasling, J., "Production of the Antimalarial Drug Precursor Artemisinic Acid in Engineered Yeast", in: *Nature* 440, 2006, pp. 940–943.

6 Bond-Watts, B. B., Bellerose, R. J. and Chang, M. C., "Enzyme Mechanism as a Kinetic Control Element for Designing Synthetic Biofuel Pathways", in: *Nature Chemical Biology* 7, 2011, pp. 222–227.

7 Anderson, J. C., Clarke, E. J., Arkin, P. A. and Voigt, C. A., "Environmentally Controlled Invasion of Cancer Cells by Engineered Bacteria", in: *Journal of Molecular Biology*, 355, 2006, pp. 619–627.

8 E.g. Elowitz M. B. and Leibler, S., "A Synthetic Oscillatory Network of Transcriptional Regulators", in: *Nature* 403, 6767, 2000, pp. 335–358; Gardner, T. S., Cantor, C. R. and Collins, J. J., "Construction of a Toggle Switch in Escherichia coli", in: *Nature* 403, 6767, 2000, pp. 339–342.

began discussions about the appropriateness of transferring concepts from physics to biology already in the mid-1990s. These discussions lead to programmatic articles such as "From molecular to modular cell biology" published in 1999 by Leland Hartwell, John Hopfield, Stanislas Leibler and Andrew Murray.[9] All four authors, two of whom are physicists (John Hopfield and Stanislas Leibler) and the other two biologists (Leland Hartwell and Andrew Murray), have made important contributions in their respective fields of research. In this article, the four authors argue for turning away from the prevailing reductionist approaches in molecular biology that "reduce biological phenomena to the behavior of molecules".[10] According to the authors, these approaches fail to take into consideration that biology-specific functions cannot be attributed to one molecule, but that "[…] most biological functions arise from the interaction among many components".[11] To describe biological functions, they go on to claim, "we need a vocabulary that contains concepts such as amplification, adaptation, robustness, insulation, error correction, and coincidence detection".[12]

To be sure, Hartwell et al.[13] paint a too reductionist picture of molecular biology and they seem to ignore early attempts to apply engineering concepts to biology – often side-by-side with concepts adapted from physics.[14] But the key point is that Hartwell et al. argue against the use of concepts taken from physics when considering biology, and instead suggest plundering the engineering lexicon. Analogies to engineered artifacts were considered appropriate as such items are typically constructed to fulfill a certain *function* – like the parts of biological organisms. This stance helped to create a collective identity for physicists entering into synthetic biology and shape the research practice of this emerging research field – a field that was attributed with a, somewhat misleading, radical novelty. However, a closer look at the development of synthetic biology reveals that it was not long before researchers began to question the validity of these engineering concepts, and subtly the meanings of the concepts began to change when applied to the design, manipulation, and exploration of synthetic biological systems.

From a philosophical perspective, it can be argued that the synthetic biologists who undertook a basic science approach did not adopt the engineering concepts and vocabulary uncritically: they actually used the genetic circuits they engineered to study, apart from the fundamental organization of biological systems, also the engineering concepts used in this endeavor. Thus there is an interesting *reflexive*

9 Hartwell, H. L., Hopfield, J. J., Leibler, S. and Murray, W. A., "From Molecular to Modular Cell Biology", in: *Nature* 402, 1999, C47–C52.

10 Hartwell, H. L., Hopfield, J. J., Leibler, S. and Murray, W. A., "From Molecular to Modular Cell Biology", in: *Nature* 402, 1999, C47.

11 *Ibid*.

12 *Ibid.*

13 *Ibid.*

14 Jacob, F., and Monod, J., "Genetic Regulatory Mechanisms in the Synthesis of Proteins", in: *Journal of Molecular Biology* 3, 1961, pp. 318–356.

twist to this endeavor, which is enabled by a new type of model – the synthetic model – developed in this field, *and* the characteristic way in which it is used. Synthetic models are typically triangulated in a combinatorial fashion with mathematical models and experiments on model organisms. In the following we discuss how the practice of combinatorial modeling has lead scientists to discover important differences between the control mechanisms of biological and engineered things.

2.2 Providing control in engineered and biological systems

Control is of central importance in engineered as well as in biological systems. However, already early on it was discovered that there are fundamental differences between controlling the behavior of biological systems and that of engineered artificial systems. Engineered systems typically rely on autonomous control mechanisms. A thermostat is a good example. In this case the room temperature (input) is measured, compared to a reference temperature (output), and in the next step the heater is changed in such a way that the room temperature is adjusted to the reference temperature. The biological solution is more elegant and makes use of internal oscillating cycles that interact and harmonize the behavior of the parts of biological organisms by coupled oscillations. Biological systems need cyclic organization, since they use the matter and energy of their environments to reconstruct and organize themselves.[15] In this biological systems differ crucially from artificial engineered systems – a point addressed by Brian Goodwin in 1960s. Goodwin was an early mathematical modeler of oscillatory feedback mechanisms and he proposed the first model of a genetic oscillator, showing that regulatory interactions among genes allowed periodic fluctuations to occur. Goodwin contrasted the behavior of genetic oscillators with engineered control systems writing: "The appearance of such oscillations is very common in feedback control systems. Engineers call them parasitic oscillations because they use up a lot of energy. They are usually regarded as undesirable and the control system is nearly always designed, if possible, to eliminate them".[16] Thus decades before the emergence of synthetic biology, it was already clear that biological organisms organize their behavior differently than the engineered artefacts.

Goodwin's model and its extensions have been used as basic templates for other models of oscillatory behavior, including the circadian clock (see Bechtel this volume). Instead of one clock it actually consists of a large orchestra of "clocks"

15 To which extent biological organisms gain control over their functioning by self-organization arising from interacting oscillations is an open question. Living systems do also rely on such decoupled controllers as genes (see Bechtel, W. and Abrahamsen, A., "Complex Biological Mechanisms: Cyclic, Oscillatory, and Autonomous", in: C. A. Hooker (Ed.), *Philosophy of Complex Systems. Handbook of the Philosophy of Science*, vol. 10. Oxford: Elsevier 2011, pp. 257–285, for an excellent discussion on the role of different oscillations in biological systems).

16 Goodwin, B., *Temporal Organization in Cells.* London: Academic Press 1963, p. 5.

that on the basis of oscillations on a molecular level synchronize the functions of the organs in a biological organism.[17] Although in comparison to circadian clocks the humanly engineered control systems, such as thermostats, appear rather simple, they are still thought to have something important in common: both make use of feedback mechanisms. One of the most basic assumptions in the modeling of control in biological systems is that they make use of feedback mechanisms. Such feedback mechanisms are typically modeled using non-linear equations, which give rise to oscillations. Yet up until recently, researchers have been uncertain whether the kinds of feedback systems depicted by the various mathematical models proposed are really realizable in biological systems. Namely, that the well-established ways of mathematically creating feedback systems used by physicists[18] may not represent the way naturally evolved organisms organize themselves. But with the advent of synthetic biology, synthetic models could be created and then it was possible to demonstrate that feedback mechanisms in biological systems can indeed lead to the kind of oscillatory behavior exhibited by circadian clocks.

2.3 Synthetic models and the combinatorial strategy

One of the defining strategies of the basic science oriented approach is the combinatorial use of mathematical models, experiments on model organisms – and synthetic models. The basic idea of this combinatorial modeling strategy is shown in Figure 1, which is taken from a review article on synthetic biology by Sprinzak and Elowitz.[19] As the upper part (a) of the diagram suggests, in combinatorial modeling the results gained from the three different epistemic activities inform each other.

17 See e.g. Bechtel, W. and Abrahamsen, A., "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science", in: *Studies in History and Philosophy of Science* 41, 2010, pp. 321–333; Bechtel, W. and Abrahamsen, A., "Complex Biological Mechanisms: Cyclic, Oscillatory, and Autonomous", in: C. A. Hooker (Ed.), *Philosophy of Complex Systems. Handbook of the Philosophy of Science*, vol. 10. Oxford: Elsevier 2011, pp. 257–285.

18 See e.g. Strogatz, S., *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Cambridge (Mass.): Perseus Books, 1994.

19 Sprinzak, D. and Elowitz, M. B., "Reconstruction of Genetic Circuits", in: *Nature* 438, 7067, 2005, pp. 443–438.

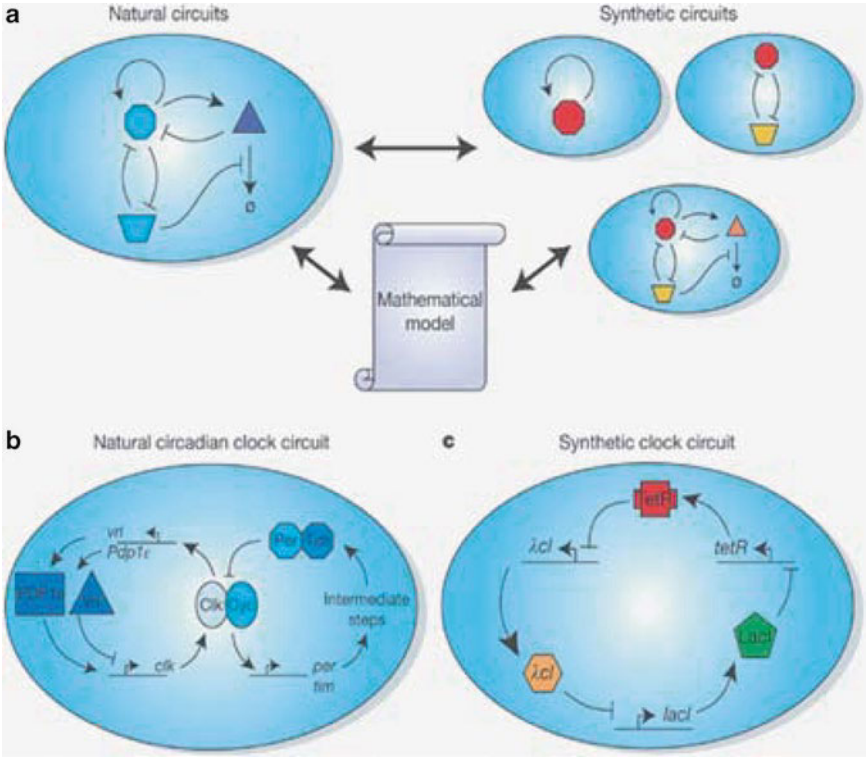


Figure 1. Combinatorial modeling according to Sprinzak and Elowitz (2005).

Why do researches make use of such a combinatorial modeling strategy in studying the organizational principles in biology? A clue can be found from the lower part (b) of the diagram. The left hand side of the diagram depicts our present understanding of the "natural gene regulatory circuit" of the circadian clock of *Drosophila* (fruit fly) consisting of interacting genes and proteins and the right hand side a synthetic model of the circadian clock, the *Repressilator*, to be introduced in the next section. The diagram indicates the two main differences between the natural and the synthetic system:

1. The natural system exhibits a much higher degree of complexity than the synthetic system.
2. The synthetic circuit has been designed by using different genes and proteins.

Consequently, synthetic models have the advantage of being less complex than model organisms. On the other hand, in comparison with mathematical models they are of the same materiality as biological systems (although the *Repressilator* was constructed from different genetic material than the naturally occurring circadian clocks, a point to which we will return below). This fact of being of the

same materiality as natural systems is crucial for the epistemic value of synthetic modeling. Roughly, it means that synthetic models are expected to work in the same way as biological systems. This very materiality of synthetic models has led researchers to discover new features of the functioning of biological systems, features that were not anticipated by mathematical modeling, or experimentation with model organisms.

2.4 The Repressilator *and the emergence of the functional meaning of noise*

The *Repressilator* is one of the first and most famous synthetic models. It is an oscillatory genetic network, which was introduced in 2000 by Michael Elowitz and Stanislas Leibler.[20] The first step in constructing the *Repressilator* consisted in designing a mathematical model, which was used to explore the known basic biochemical parameters and their interactions. Next, having constructed a mathematical model of a gene regulatory network Elowitz and Leibler performed computer simulations on the basis of it. They showed that there were two possible types of solutions: "The system may converge toward a stable steady state, or the steady state may become unstable, leading to sustained limit-cycle oscillations".[21] Furthermore, the numerical analysis of the model gave insights into the experimental parameters relevant for constructing the synthetic model and helped in choosing the three genes used in the design of the network.

The structure of the *Repressilator* is depicted in the following diagram:

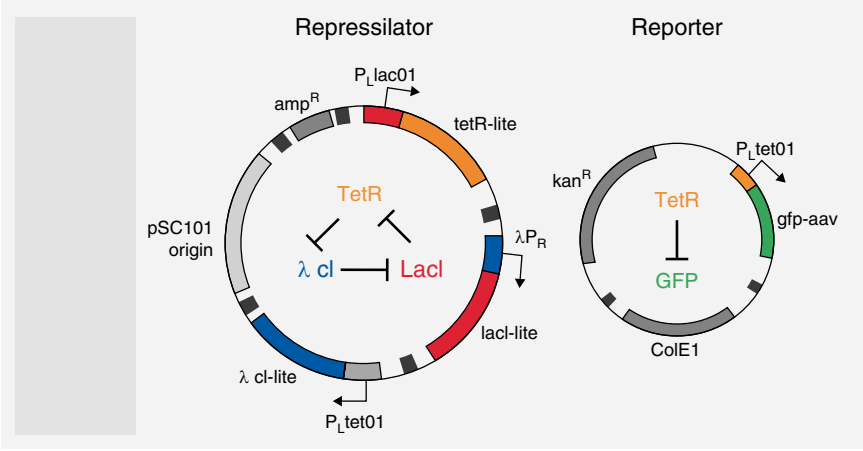


Figure 2. The main components of the *Repressilator* (left hand side) and the *Reporter* (right hand side) (Elowitz and Leibler 2000, p. 336).

20 Elowitz M. B. and Leibler, S., "A Synthetic Oscillatory Network of Transcriptional Regulators", in: *Nature* 403, 6767, 2000, pp. 335–358.

21 *Ibid.,* p. 336.

In the diagram the synthetic genetic regulatory network, the *Repressilator,* is shown on the left hand side and it consists of two parts. The outer part is an illustration of the plasmid constructed by Elowitz and Leibler. The plasmid is an extra-chromosomal DNA molecule integrating the three genes of the *Repressilator*. Plasmids occur naturally in bacteria. In the state of competence, bacteria are able to take up extra chromosomal DNA from the environment. In the case of the *Repressilator,* this property allowed the integration of the specific designed plasmid into *E. coli* bacteria. The inner part of the illustration represents the dynamics between the three genes, *TetR*, *LacI* and *λcI*. The three genes are connected by a negative feedback loop. The right hand side of the diagram shows the *Reporter* consisting of a gene expressing a green fluorescent protein (GFP), which is fused to one of the three genes of the *Repressilator*. The GFP oscillations in the protein level made visible the behavior of transformed cells allowing researchers to study them over time by using fluorescence microscopy.

The construction of the *Repressilator* was enabled by the development of new methods and technologies, such as the construction of plasmids and Polymerase Chain Reactions (PCR). On the other hand, the construction of synthetic models has so far been limited to simple networks such as the *Repressilator* whose construction components (and their number) had to be chosen in view of what would be optimal for the behavior under study.[22] This means that such networks need not be part of any naturally occurring system. For example the genes used in the *Repressilator* do not occur in such a combination in any biological system but were chosen and tuned on the basis of the simulations of the underlying mathematical model and other background knowledge in such a way that the resulting mechanism would allow for (stable) oscillations.

Summing up: above we have described how with the formation of synthetic biology a new tool was introduced into the research on biological organization: the construction of novel engineered genetic networks specially designed for answering certain kinds of theoretical questions. Mathematical models were unable to settle the question of whether the various network designs proposed, e.g. in the context of circadian clock research, could actually work in biological organisms. This problem was aggravated by the fact that the model templates, methods and concepts used were not originally devised with biological organisms in mind. Neither could this problem of the generality and foreignness of the theoretical tools used be conclusively settled by experimentation since the work with model organisms had to deal with the immense complexity of even such simple model organisms as *E. coli.* Moreover, experimentation relies on mathematical modeling in the interpretation of experimental results. Thus even though empirical research has progressed considerably over recent decades with respect to studying the genes and proteins involved in the circadian clock phenomena, for example, the

22 In the case of the *Repressilator* the order in which the genes are connected to each other, turned out to be crucial, too.

results are often inconclusive. Synthetic models, like the *Repressilator* are partly able to fill the gap between mathematical modeling and experimentation on model organisms by offering a tool for identifying possible network design principles, and showing whether they might be realizable in biological organisms. Moreover, by implementing the synthetic genetic network into a cell it is exposed to some further constraints of natural biological systems, thus providing insight into the modularity of the circadian mechanism. Interestingly, the *Repressilator* sparked a new line of research as a direct result of its limited success. In contrast to the mathematical model underlying it, the *Repressilator* did not show the expected behavior: regular oscillations. Instead, the oscillations turned out to be noisy. Computer simulations suggested that stochastic fluctuations could be the cause of this noisy behavior. This led researchers to explore the meaning of noise in the context of biology. An exploration that in itself highlighted further differences between engineered artefacts and biological systems. Whereas in engineering noise is usually regarded as a disturbance, the recent research in synthetic biology indicates that in biological organisms noise also plays a functional role. Biological systems appear to make good use of noise in diverse processes, including development,[23] differentiation (e.g. genetic competence[24]), and evolution.[25] Apart from internal noise, there remained the possibility that the noisy behavior could also have been caused by external noise coming from the cell environment. This in turn means that the *Repressilator* was probably not so modular as it was supposed to be, that is, it did not form as isolated a module in its host system as was expected. Indeed, apart from noise, modularity is another engineering concept whose limits have been questioned by recent research in synthetic biology.

3. The second wave of synthetic biology: Aiming for integration

3.1 Investigating the modularity assumption

Modular organization is among the most basic and important assumptions of synthetic biology, but also one of the most contested ones. Since its beginning synthetic biology has faced the following dilemma regarding the assumption of modular organization: on the one hand, synthetic biology relies on the assumption of modular organization in view of its aim to design autonomous modules of

23 Neildez-Nguyen, T. M. A., Parisot, A., Vignal, C., Rameau, P., Stockholm, D., Picot, J., Allo, V., Le Bec, C., Laplace, C. and Paldi, A., "Epigenetic Gene Expression Noise and Phenotypic Diversification of Clonal Cell Populations", in: *Differentiation* 76, 1, 2008, pp. 33–40.

24 Çagatay, T., Turcotte. M., Elowitz M. B., Garcia-Ojalvo. J. and Süel, G. M., "Architecture-Dependent Noise Discriminates Functionally Analogous Differentiation Circuits", in**:** *Cell* 139, 3, 2009, pp. 512–522.

25 Eldar, A. and Elowitz, M. B., "Functional Roles for Noise in Genetic Circuits", in: *Nature* 467, 2010, pp. 167–173.

interacting components that would give rise to a specific function/behavior. On the other hand, each synthetic biological system also functions as a test to which extent the assumption of the modular organization is justified.

Looking at more recent developments in synthetic biology it seems that synthetic biologists, forced by the insights they have gained from designing and constructing synthetic systems, have begun to reconsider the assumption of modularity. They have left behind the *strictly* modular organization and allowed for some interaction between the components of a module and the other constituent parts of the cell in which it is embedded. This more close integration of synthetic systems with the host cell means a loss of control over the performance of the synthetic system but it also opens up new possibilities for the design of synthetic systems. This situation is very similar to the case of noise. Noise in biological systems also has two sides: from the engineering perspective it means losing partial control over the performance of a synthetic system, but, on the other hand, noise also has a functional component that improves the performances of an organism. Thus for synthetic biology the critical point is how to make use of noise in the design and engineering of synthetic systems, or in the case of modular organization, how to integrate the components of synthetic systems with those of the host cell to support the performance of the synthetic system. Nagarajan Nandagopal and Michael Elowitz[26] put forward one possible strategy. The two authors explicate what they mean by integration on the systems level by referring to a work by Stricker et al.[27] on a transcriptional oscillator. The design of this oscillator is even simpler than that of the *Repressilator* – it just consists of two genes: an activator and a repressor. The expression of either gene can be enhanced by the activator protein and blocked by the repressor protein. Both proteins function as transcription factors for both genes. Concerning the dynamic of their model system, Stricker et al. made the interesting observation that unintended interactions of the synthetic system with the host cell actually improved the oscillatory behavior of the system by making the oscillations more precise.

Consequently, and in contrast with the traditional aim of designing isolated modules, the interactions between synthetic systems and the host cell need not always be a bad thing, but could be advantageous as well. Having pointed this out, Nandagopal and Elowitz proceed to call for synthetic systems "that integrate more closely with endogenous cellular processes".[28] With this step, they suggest, the field would move away from its original aim of designing "autonomous genetic circuits that could function as independently as possible from endogenous

26 Nandagopal, N. and Elowitz, M. B., "Synthetic Biology: Integrated Gene Circuits", in: *Science* 333, 2011, pp. 1244–1248.

27 Stricker, J., Cookson, S., Bennet, M. R., Mather, W. H., Tsimring, L. S. and Hasty, J., "A Fast, Robust and Tunable Synthetic Gene Oscillator", in: *Nature* 456, 2008, pp. 516-519.

28 Nandagopal, N. and Elowitz, M. B., "Synthetic Biology: Integrated Gene Circuits", in: *Science* 333, 2011, pp. 1244–1248.

cellular circuits or even functionally replace endogenous circuits".[29] Nandagopal and Elowitz use a three-partite picture (Figure 3) to depict what they think will be one of the big changes in the practice of synthetic biology: "Future progress will require work across a range of synthetic levels, from rewiring to building autonomous and integrated circuits de novo".[30]

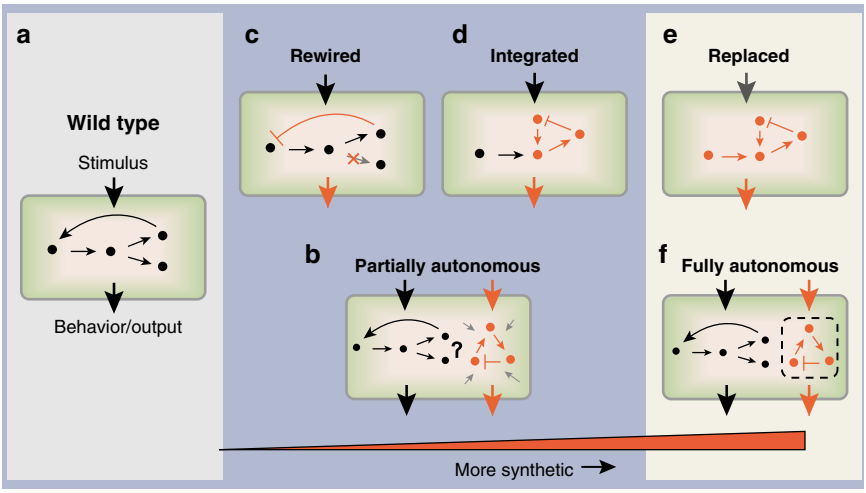


Figure 3. The continuum of synthetic biology (Nandagopal and Elowitz 2011, p. 1244).

In the diagram depicted in Figure 3 Elowitz and Nandagopal introduce what they call the "continuum of synthetic biology". In this continuum one moves from the wild type towards fully autonomous synthetic systems increasing the degree of the synthetic part of the system. How is this increase in the synthetic part achieved? There are several options. One can follow the "traditional" approach of designing an assumedly modular genetic circuit and introducing it into the wild type. As the example of Stricker et al. nevertheless showed, unintended interactions can occur (gray arrows) that could be difficult to control. An alternative approach, propagated by Nandagopal and Elowitz, consists in first rewiring the genetic circuit in the wild type and then in a second step implementing a synthetic circuit into the rewired circuit. This rewiring of the existing genetic circuits offers, firstly, a way to explore the design principles on which the genetic circuit is based and, secondly, a possibility of using these insights to avoid unintended interactions with the host cell. As has been shown in a number of studies in which the strategy of rewiring

29 *Ibid.* p. 1244.
30 *Ibid*. p. 1244.

has been used, the actual biological design principles often are counter-intuitive.[31] Nature appears to have used solutions which differ from those of engineers.

As a consequence of the rewiring strategy the resulting engineered circuit is only partially independent. However, for the engineering purposes as high modularity as possible is usually sought because of its controllability. In order, then, to get an independent circuit that would be based on the insights gained from the exploration of the rewired circuit one would integrate the function of the rewired circuit design into an autonomous genetic circuit. This strategy allows for suppressing unwanted interactions with the host cell but also implementing interactions which support the function in question. In more general terms, the proposed strategy tries to balance the need for control and the possibility of taking advantage of the interactions with the host cell. In such a way the engineering of synthetic systems becomes increasingly inspired by biological systems – a point that has recently been stressed by several synthetic biology research programs.[32]

3.2 The call for disciplinary integration

According to the latest developments in synthetic biology, the field seems to be ready for new challenges. From a stage in which the main goal consisted in exploring the applicability of engineering principles in the context of biology, the synthetic biologists working in the basic science branch are moving forward towards more concrete applications. Or as the Ruder, Lu and Collins put it:

> The field initially arose from the combined efforts and insights of a small band of engineers, physicists, and computer scientists whose backgrounds dictated the early directions of synthetic biology. For the field to reach its full clinical potential, it must become better integrated with clinicians.[33]

Thus the above-mentioned integrational approach in the exploration of the basic design principles of biological organization is accompanied with the call for integration also on the disciplinary level.[34] In order to find novel ways and strategies for instance in medicine, synthetic biologists feel that they need the support and know-how of clinical researchers. Combining the integration efforts on these two

31 See e.g. Çagatay, T., Turcotte. M., Elowitz M. B., Garcia-Ojalvo. J. and Süel, G. M., "Architecture-Dependent Noise Discriminates Functionally Analogous Differentiation Circuits", in**:** *Cell* 139, 3, 2009, pp. 512–522.

32 See e.g. http://wyss.harvard edu/viewpage/264/a-new-model. Accessed at 5 January 2012.

33 Ruder, W. C., Lu, T. and Collins, J. J., "Synthetic Biology Moving into the Clinic", in: *Science* 333, 2011, p. 1251.

34 O'Malley and Soyer argue that systems and synthetic biology provide good examples of the various kinds of integrative pursuits taking place in contemporary science, see O'Malley, M. A. and Soyer, O. S., "The Roles of Integration in Molecular Systems Biology", in: *Studies in History and Philosophy of Biological and Biomedical Sciences*, 2011, pp. 58-68.

fronts is an ambitious aim, but synthetic biologists find in such fields as clinical research, a lot of potential for the application of their specific engineering approach. The long list of possible clinical applications includes the treatment of infectious diseases and cancer, as well as vaccine development, microbiome engineering, cell therapy, and regenerative medicine.[35]

For instance, in cancer research synthetic biology could design and produce special bacteria, which would be able to identify and kill cancer cells. The possibility of targeting only cancer cells would have the advantage of avoiding the side effects of traditional cancer therapies, such as the damage of healthy tissue. Ruder, Lu and Collins[36] argue that for these developments to take off, synthetic biologists have to integrate their research and engineering efforts into the research done in clinical labs. Synthetic biologists believe that the experiences they have accumulated in the manipulation of synthetic biological systems empower them to offer clinical practice biologically inspired and hopefully also practically implementable solutions.

4. Conclusion

Above we have argued that in contrast to the popular image of synthetic biology as a discipline attempting to force biological systems into an engineering mold, the exploration of the differences between engineering and biology has been one of the central foci of the basic science approach to synthetic biology. The materiality of synthetic biological systems and the possibility of directly manipulating biological components has provided many valuable insights into biological organization as well as pointed towards the limitations of any single-minded engineering approach. What seems in our opinion to be too easily glossed over by the critics of synthetic biology is the fact that in engineering synthetic biological things synthetic biologists are at the same time also exploring the assumptions on which this endeavor is built. This reflexive element in their endeavor has, in a relatively short time, made synthetic biologists aware of some fundamental differences between the functioning of engineered artifacts and biological organisms. As the recent work on the concepts of noise and modularity show, synthetic biology is in the process of becoming more "biology inspired". These new insights do not make the engineering of synthetic biological systems an easier task – rather, they increase our awareness of the difficulties and challenges to be encountered.

35 Ruder et al., *ibid.* p. 1249.
36 *Ibid.*

Tarja Knuuttila
University of Helsinki
Fabianinkatu 24 (P. O. Box 4)
00014, Helsinki
Finland
tarja.knuuttila@helsinki.fi

Andrea Loettgers
California Institute of Technology
1200 E. California Blvd., MC 114-96
Pasadena, CA 91125
USA
loettger@caltech.edu

Anders Strand and Gry Oftedal

Causation and Counterfactual Dependence in Robust Biological Systems[1]

Abstract

In many biological experiments, due to gene-redundancy or distributed backup mechanisms, there are no visible effects on the functionality of the organism when a gene is knocked out or down. In such cases there is apparently no counterfactual dependence between the gene and the phenotype in question, although intuitively the gene is causally relevant. Due to relativity of causal relations to causal models, we suggest that such cases can be handled by changing the resolution of the causal model that represents the system. By *decreasing* the resolution of our causal model, counterfactual dependencies can be established at a higher level of abstraction. By *increasing* the resolution, stepwise causal dependencies of the right kind can serve as a sufficient condition for causal relevance. Finally, we discuss how introducing a temporal dimension in causal models can account for causation in cases of non-modular systems dynamics.

1. Introduction

Counterfactual dependence accounts of causation have several problems accounting for causation in complex biological systems (Mitchell 2009, Strand and Oftedal 2009). Often perturbations on such systems do not have any clear-cut phenotypic effects, and consequently there is no direct counterfactual dependence between the cause candidate intervened on and the effect considered. For example, many gene knockouts and knockdowns have no detectable effect on relevant functionality, even though the genes in question are considered causally relevant in non-perturbed systems (Shastry 1994, Wagner 2005).

Two different mechanisms give rise to such stability: (1) gene redundancy; the workings of backup-genes explain the lack of counterfactual dependence between the effect and the preempting cause, and (2) distributed robustness; the system readjusts functional dependencies among other parts of the system rather than invoking backup genes. The latter cases challenge not only the necessity of coun-

1 Thanks to Henrik Forssell, Sara Green, Carsten Hansen, Heine Holmen, Veli-Pekka Parkkinen, the audience at the ESF Philosophy of Systems Biology Workshop, Aarhus University, and the participants at the colloquium for analytic philosophy, Aarhus University, for helpful comments and suggestions.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_15, © Springer Science+Business Media Dordrecht 2013

terfactual dependence for causation, but also our thinking about truth conditions for the relevant counterfactuals. We suggest that such cases can be handled by a counterfactual dependence account of causation by changing the resolution of the causal model.

There is a related problem concerning non-modularity of complex systems. Modularity is by some seen as a requirement for making adequate causal inferences (Woodward 2003). The intuitive idea is that different mechanisms composing a system are separable and in principle independently disruptable (Hausman and Woodward 1999). However, research indicates that compensatory changes in response to disruptions in biological systems can change functional relations between relevant variables and thereby violate modularity.

In the following we first introduce the core elements of a counterfactual dependence based philosophical analysis of causation. Then we present gene redundancy and distributed robustness using biological examples and argue that changing representational resolution helps understand causal dependence in these cases. Finally, we discuss how introducing a temporal dimension in causal models gives a grip on non-modular systems dynamics.

2. Sketch of an analysis of causation

Causes typically make a difference to their effects, and many philosophers argue that this idea should be at the core of the philosophical analysis of causation (e.g. Lewis 1973, 2004, Woodward 2003, Menzies 2004). We agree and suggest the following general definition of causation, where X and Y are variables and M is a causal model, i.e. a set of variables and functional relations between them:

> ***Causal Relevance****:* X is a cause of Y relative to M if and only if there is a change of X that would result in a change of Y when we hold some subset of variables (allowing this set to be empty) in M fixed at some values.

This definition is in line with other well discussed difference-making accounts of causation viewing causal relata as variables (e.g. Menzies and Woodward). Such views capture the idea that causal relations are exploitable for purposes of manipulation and control.

The requirements of *some* subset of variables being held fixed at *some* values are chosen with care. The main idea is that this definition states causal relevance in the broadest sense, and that different explications of the relevant subset of variables and their relevant values give different kinds of causal relevance. Letting the subset be empty, for example, gives the notion of a total cause:

> ***Total Cause***: X is a total cause of Y relative to M if and only if there is a change of X that would result in a change of Y.

Also the notion of a direct cause comes out as a special case:

> ***Direct cause***: X is a direct cause of Y relative to M if and only if there is a change of X that would result in a change of Y when we hold all other variables in M fixed at some values.

Causal paths are understood as chains of relations of direct causation. Using this idea, we can cash out the notion of a contributing cause:

> ***Contributing Cause***: X is a contributing cause of Y relative to M if and only if there is a change of X that would result in a change of Y when we hold all variables in M not on the relevant path from X to Y fixed at some values.

'Relevant path' is a placeholder for the chain of direct causal relations between X and Y over which we are checking for mediated causal relevance among X and Y. The existence of a mediating chain of direct causal relationships is not itself sufficient for causal relevance due to counterexamples to transitivity. Furthermore, actual causation can be specified in terms of all variables not on the relevant path between X and Y being held fixed at their actual values.

These distinctions, which mirror Woodward's 2003 distinctions, are not exhaustive. We add two additional notions here, tentatively called *Restricted Causal Relevance* and *Dynamic Causal Relevance* (Section 5). The idea behind restricted causal relevance is to capture causal understanding often implicit in actual scientific practice, where variables not tested for causal relevance are held fixed at assumed normal or expected values. This is *restricted* causal relevance because it requires counterfactual dependence under a limited range of values of the variables held fixed.

> ***Restricted Causal Relevance****:* X is a restricted cause of Y relative to M if and only if there is a change of X that would result in a change of Y when we hold all variables in M not on the relevant path from X to Y fixed at their normal values.

There will be a variety of different notions of restricted causal relevance. The one stated here is analogous to contributing cause, and should be sufficient to illustrate the core idea.

Counterexamples to the claim that counterfactual dependence is necessary for causation feature a redundancy of cause candidates: preemption (a cause preempts a backup cause), overdetermination (two individually sufficient causes), and trumping (a cause trumps another cause candidate). Moreover, distributed robust-

ness found in biological systems presents yet another counterexample, one that has not received sufficient attention in the philosophical literature, but has some interesting and perhaps surprising philosophical consequences.

It is important to be aware that preemption cases are only problematic for the more demanding varieties of causal relevance. For a standard case of preemption to arise in the first place, both cause candidates C_1 and C_2 must be causally relevant to E in the broad sense of causal relevance (see Section 4). When philosophers ask which of C_1 and C_2 are actual or restricted causes of E problems occur, because asymmetry intuitively should arise for these more demanding notions. It is only if C_2 is causally relevant in the broad sense that it can be a preempted backup cause in actual or normal circumstances where it is not an actual or restricted cause itself. Allowing for a proper description of type-level preemption cases is the main role of the notion of restricted causal relevance in this paper.

3. Genetic redundancy and distributed robustness

Gene knockout and knockdown experiments investigate the functioning of genes by effectively deleting or silencing specific genes (e.g. Xie et al. 2005). Hypothesis-driven experiments of this sort often involve causal reasoning of the form that if the procedures make a difference to a particular phenotypic trait, then the gene in question is causally relevant for that trait. However, due to system robustness, very often gene perturbations do not have any apparent effect on the functionality of the system at hand (Shastry 1994, Wagner 2005).

Robustness is a ubiquitous property of living systems and allows systems to maintain their biological functioning despite perturbations (Kitano 2004, 826). Mutational robustness can be described as functional stability against genetic perturbations (Strand and Oftedal 2009), and two types are recognized in the literature; genetic redundancy and distributed robustness (Wagner 2005). Genetic redundancy involves multiple copies of a gene or genes with similar functionality (so-called duplicate genes) that can take the role of the perturbed gene. Distributed robustness is more complex and involves organizational changes of multiple causal pathways in such a way that the system manages to compensate for the genetic disturbance (Hanada et al. 2011).

Genetic redundancy was investigated in Kuznicki et al. (2000), where duplicate genes were found to contribute to robustness in the nematode *C. elegans*. GLH proteins (GermLine RNA Helicases) are constitutive components of the nematode P granules. These granules are distinctive bodies in the germ cells found to have roles in the specification and differentiation of germ line cells. The genes associated with the proteins GLH-1 and GLH-4 belong to the multi-gene GLH family in *C. elegans*, and the GLHs are considered important in the development of egg cells (oogenesis). Still, no effect on oogenesis could be detected either from

a RNAi knockdown of the gene associated with GLH-1 or with GLH-4. However, the combinatorial knockdown of both the GLH-1 and GLH-4 genes resulted in 97% sterility due to lack of egg cells and defective sperm. The results indicate that GLH-1 and GLH-4 are duplicate genes that can compensate for each other when one or the other is lacking. We return to this example in Section 4.

Distributed robustness was investigated in Edwards and Palsson 2000a and 2000b where chemical reactions in *E. coli* were perturbed. Rather than knocking out or knocking down genes, chemical reactions in the process of glycolysis (the metabolic process of converting glucose into pyrovate and thereby produce the energy rich compounds ATP and NADPH) were blocked one by one to find whether any of the reactions were essential to cell growth. Only seven of the 48 reactions were found to be essential, and of the 41 remaining, 32 reduced cell growth by less than 5%, and only nine reduced cell growth with more than 5%. For example, the blocking of the enzyme G6PD (glucose-6-phosohate dehydrogenase) resulted in growth at almost normal levels. However, the elimination of this reaction had major systemic consequences (Wagner 2005). Instead of producing two-thirds of the cell's NADPH (a coenzyme needed in lipid and nucleic acid synthesis) by the pentose phosphate pathway, more NADH was produced through a different path, the tricarboxylic acid cycle, and this NADH was then transformed into NADPH via a highly increased flux through what is called the transhydrogenase reaction. In other words, practically all the NADPH needed for upholding normal cell growth was still produced, but through different pathways. We return to this example in Section 5.

4. Dealing with redundancy

When a duplicate gene takes the role of a silenced gene, there is typically no phenotypic change that indicates causal relevance of the silenced gene[2]. Consider a standard case of late preemption (Figure 1).

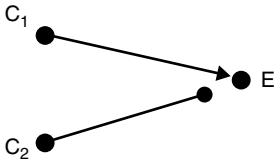


Figure 1

2 Certain fine-grained redescriptions of the phenotypic effect may still reveal dependencies. This could be cashed out in terms of a change of causal model or different restrictions on the values of the variables.

According to our definition of type level causal relevance, both cause candidates are causally relevant, but only C_1 is actual cause. Type level preemption cases can be formulated in terms of restricted causal relevance. C_1 and C_2 may both be broadly causally relevant, while both should not be restricted causes. Suppose that normal values of C_1 and C_2 are that both are present. Then we should be able to say that in normal cases, C_1 is a restricted cause, while C_2 is not, even though C_2 would be a cause in abnormal knockout cases where C_1 is not present.

One might think that this asymmetry could be accounted for in terms of specificity (Woodward 2010) or fine-tuned influence (Lewis 2004, 92). The idea is that there is an asymmetry between the preempting cause and the preempted backup, because the relation between the preempting cause and the effect is such that one can make minor changes to the cause that are followed by minor changes in the effect, while there is no analogue for the preempted backup. Intervening to slightly alter the preempted backup will not change the effect at all. This strategy is promising, and can cover several cases, but it does not work in full generality. In particular, it does not handle cases of threshold causation, where fine-tuning of the preempting cause either changes the effect from occurring to not occurring, or makes no difference at all.

On our definitions, like on Woodward's (2003) and Menzies' (2004), causal relations obtain relative to a causal model. On this background, we see how causal dependencies can be masked and/or revealed by changing the resolution of the causal model, for example by invoking a more coarse-grained model. Consider a simplified gene-redundancy scenario (Figure 2). Binary variables v_1, v_2, and v_3 represent the presence or absence of three functionally similar genes. Consider another representation involving only one binary variable, v_4, that takes the value present when at least one of v_1, v_2, or v_3, take the value present, and the value absent when all of v_1, v_2, and v_3 take the value absent.

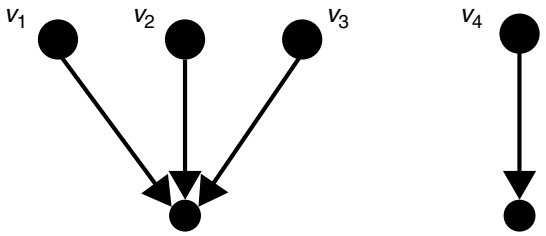


Figure 2

Interventions changing the value of v_4 will directly affect the effect variable in this setup, and the counterfactual dependence of the effect on v_4 is straightforward. Applied to the example of redundancy in the previous section, this corresponds to

considering the disjunction of the two mutually compensating genes GLH-1 and GLH-4 as the relevant variable in relation to germ cell formation and not the presence of the individual genes. Whether higher-level representations like these have, or should have, realist interpretations in terms of e.g. modally robust entities is a further question (see Strand and Oftedal 2009). Alternatively, one may introduce a fine-grained effect-variable in order to reveal fine-tunable causal influence.

5. Dealing with distributed robustness

In systems exhibiting distributed robustness, organization and causal paths can be rewired under perturbations. Such systems can retain their biological functions by changing their causal structure compared to the structure they would have had in the absence of that perturbation. Systems with such behavior can be non-modular in the sense that the intervention on one causal factor, for example a gene, changes causal relations between other factors in the system.[3] *Prima facie*, the perturbed gene is a cause in the normal case even if there is a distributed back-up mechanism at play in the perturbed case. The causal analysis should account for the gene being causally linked to the relevant phenotypic trait in the normal case, even in the absence of the right kind of direct counterfactual dependence.

We consider two options. One is to decrease the representational resolution by abstracting away from details, and thereby in effect treating systems with distributed robustness as modules that are not internally modular. Interventions on such systems will be radical; wipe out the whole module. The other is to increase the representational resolution, in the sense that one zooms in on the relevant gene and the causal paths leading from that gene to the effect in question. This is done by introducing causal intermediaries and tracking stepwise causal dependencies. If one could establish stepwise counterfactual dependence, one could for example take the ancestral relation of counterfactual dependence, and thus establish the causal status among distant nodes that are not related directly by counterfactual dependence.[4] We elaborate on the second option in the following.

First, consider a relatively simple abstract case of distributed robustness (Figure 3).

3 E.g. Woodward (2003) and Cartwright (2001) discuss modularity.

4 The idea in David Lewis' original account is that whenever you have a chain c_1, c_2, …, c_n and each c_m and c_{m+1} are related by counterfactual dependence, any two distant elements in that chain is causally related by definition. Taking the ancestral was crucial to secure transitivity for Lewis, but transitivity is questioned in the contemporary debate (e.g. Woodward 2003, Paul 2004).

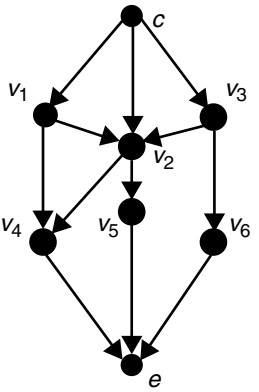


Figure 3

Imagine a knockout on v_1 that changes the functional dependencies between some of the other parts. The point is not exactly what changes are being induced, rather that such changes indeed occur. In the redundancy case, no such changes occur, it is simply a backup gene performing the function of the knocked out gene.

Now, consider the path v_1, v_2, v_5, to e. If this is a causal path, there will be causal interventions on v_1 that change the value of v_2, causal interventions on v_2 that change the value of v_5, and causal interventions on v_5 that change the value of e. However, the changes in v_2 brought about by changes of v_1 may not be such that they induce changes in v_5. Rather, it may be that the interventions on v_2 that do change v_5 cannot be induced by intervening on v_1. In such a case, there will be no direct counterfactual dependence, but there will still be a path between v_1 and e. The question is whether v_1 qualify as a cause of e? If we straightforwardly take causation to be the ancestral of counterfactual dependence, v_1 will qualify as a cause of e. However, this is too weak and deems some non-causal relations causal.

On the other hand, one might think that causal relevance between v_1 and e is mediated via a causal path if and only if there is a causal intervention on v_1 that changes the value of e (Woodward 2003 requires this). This, however, is too strong. In a case of distributed robustness, changes of v_1 that brings about certain changes in v_2 might also trigger distributed backup mechanisms that affect whether v_2 and/or v_5 can bring about changes in e. If the system is non-modular, it will be impossible to control for such backups by holding other variables fixed. We need to find some middle ground between taking the ancestral which is too weak, and requiring direct counterfactual dependence which is too strong.

Here is a tentative account. It is sufficient for causal relevance that there are changes of v_1 that result in changes of v_2, changes within the range of changes that can be brought about on v_2 by changing v_1 that result in a range of changes in v_5, and finally the same for v_5 and e. If there is such a series of ranges of changes, then

v_1 is a cause of *e* even if there are no changes of v_1 that would result directly in changes of *e*. Let's label this the relevance requirement.

An example is the metabolic reactions in *E. coli* presented previously. The distributed robustness of these reactions makes sure that practically the same amount of the energy rich compound NADPH is produced even though the pentose phosphate pathway, which normally is considered the main source of NADPH, is blocked by knocking out the enzyme G6PD (Edwards and Palsson 2000a, 2000b). As shown in the figures below, NADPH is mainly produced through the pentose phosphate pathway when no chemical reactions are blocked. When G6PD is knocked out, however, NADPH production goes through different pathways. The tricarboxylic acid cycle produces NADH at elevated levels and this NADH is transformed into NADPH through the transhydrogenase reaction.

The following illustrations are adapted from Edwards and Palsson (2000b). Additional nodes are introduced to represent the key chemical reactions DH (dehydrogenation) and DC (decarboxylation). Black arrows represent the main causal pathways (high flux). Grey arrows represent minor causal pathways (low flux). Dashed arrows represent no-flux pathways. Figure 4 shows a causal representation of glucose metabolism in *E. coli* under normal circumstances. NADPH is mainly produced through the pentose phosphate pathway (in black).

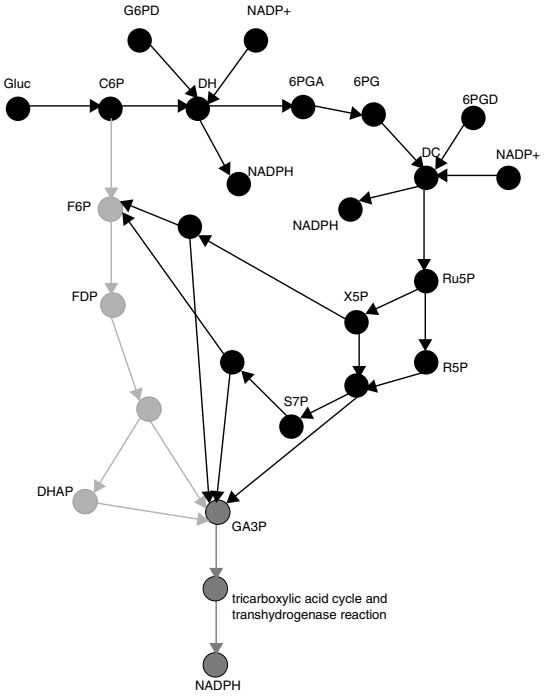


Figure 4

Figure 5 shows a causal representation of glucose metabolism in *E. coli* when the enzyme G6PD is knocked out. No NADPH is produced through the pentose phosphate pathways (now in grey). Rather, there is an increased activity in different pathways ultimately leading through the tricarboxylic acid cycle and the transhydrogenase reactions (details omitted from the figure).

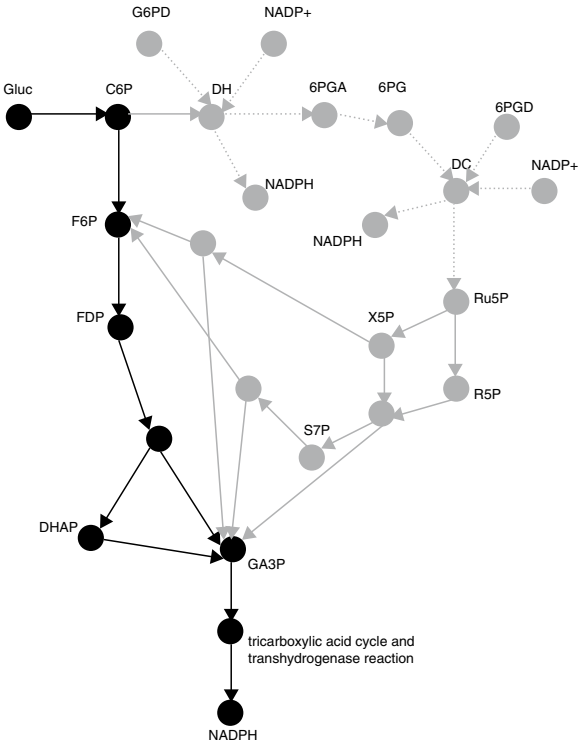


Figure 5

Even though there appears to be no counterfactual dependence between the enzyme G6PD that catalyzes parts of the pentose phosphate pathway and the total NADPH production, G6PD should still be considered causally relevant for the production of NADPH. Translating our tentative account of how causal relevance can be established into this example gives the following: It is sufficient for causal relevance that there are changes on G6PD that result in a range of changes of DH, changes within that range of changes on DH that result in a range of changes in 6PGA and NADPH, the same for 6PGA and 6PG, for 6PG and DC, and for DC and NADPH. If there is such a series of ranges of changes, then G6PD is a cause of NADPH even if there are no changes on 6PGA that would result in changes in NADPH production directly.

This account demonstrates a need to represent the effects of perturbations on the dynamic evolution of systems. Ignoring the dynamic dimension, and assuming that the systems remain fixed under perturbations, is deeply problematic when confronted with more complex system behavior. Intervening at a certain point in the dynamic evolution of a system may change the upcoming development, and when intervening at a later point we have the choice of intervening on a system that has not been disturbed, or on a *different system*, namely the one that was disturbed at an earlier point. These will be represented by different system trajectories along the dynamic dimension.

Such systems are not modular since we cannot intervene on causal paths of interest without changing other aspects of the systems. According to some philosophers (e.g. Woodward 2003), systems that are sufficiently non-modular are not causal. We think this response is too hasty. Moreover, eschewing this as a problem for dependence accounts of causation still leaves the problem of how we should understand counterfactual claims about non-modular functionally robust systems. We will therefore proceed treating it as a question about causation, trusting that it has philosophical value even if one should choose to label it otherwise.

Consider the following notion of causal relevance:

> ***Dynamic Causal Relevance***: X is a dynamic cause of Y relative to M if and only if there is a possible change of X that would result in a change of Y when we hold all variables in M fixed at some values at the time of intervention on X.

This notion allows for systemic changes over time due to earlier perturbations. We can represent the system in n-dimensional space, where n is the number of variables describing the system. The changes in the dynamic dimension can include changes of the functional relations among different parts of the system. In effect some system trajectories in this dimension will represent different systems than other trajectories. It might happen, for example in cases of distributed robustness, that the post-intervention system changes not only the values of the variables, i.e. its state, but also the functional dependencies among the variables. For such cases, we need an account that tells which counterfactual scenarios are relevant for evaluating counterfactual claims about the non-perturbed system.

Prima facie, there are two options. First, one may consider the perturbed system at a later point in its dynamic evolution, but this can mask causal relations since the system may have changed, and backups may have been triggered as a result of the perturbation. Second, one may intervene on a non-disturbed system identical to the system of interest up to the time of interest. Which choice we make can affect what relations come out as causal.

Since this is a question of when we can infer mediated causal relationships we need to get clearer on the general question of causal transitivity. The simplest general form of the standard counterexamples to transitivity requires a setup like the following:

Variables	v_1	v_2	e
Possible values	{a, b}	{1, 2, 3}	{@, $}
Possible changes	(a, b)	(1, 2)	(@, $)
	(b, a)	(1, 3)	($, @)
		(2, 1)	
		(2, 3)	
		(3, 1)	
		(3, 2)	

v_1 and e are binary variables, while v_2 can take three different values. Possible changes are represented by ordered pairs of values of the variables. In general, for a n-ary variable there will be n(n-1) possible changes of values. The counterexamples arise when the changes that can be brought about in v_2 by intervening on v_1 do not overlap with the changes in v_2 that will result in changes on e.

Woodward gives an example where a dog bites his right hand. At the type level this can be represented by a binary variable taking the values {dog bites, dog does not bite}. The bite causes him to push a button with his left hand rather than with his right hand. This intermediate cause can be represented by the triadic variable {pushes with right hand, pushes with left hand, does not push}. The pushing of the button causes a bomb to go off, represented by the binary variable {bomb explodes, bomb does not explode}. Appeal to causal intuition tells us that the bite causes the pushing, the pushing causes the explosion, but the bite does not cause the explosion. When there are no changes of v_1 that result in changes in e, v_1 is not a cause of e even though there is a chain of direct causal dependencies connecting v_1 and e. Woodward's way of dealing with the counterexamples is to deny that the ancestral of direct causal dependence is sufficient for causal relevance. To get a sufficient condition he requires interventions on v_1 that change the value of e when all variables not on the path from v_1 to e are fixed at suitable values. This latter requirement, however, is too strong.

What is needed to block the counterexamples to transitivity is a requirement of relevance and not of direct dependence. The changes brought about in v_2 by changing v_1 must be such that inducing some of *those changes* in v_2 results in changes of e. This relevance requirement accounts for the problem cases of transitivity more surgically. In particular, it leaves open the possibility that the existence of *the right kind* of causal chain is sufficient for causal relevance, even in the absence of direct dependence. In light of our earlier discussion of non-modular systems exhibiting distributed backup mechanisms, we can understand how such cases may arise. A variable can be causally relevant for an effect further downstream a certain causal path P, even if changes of that variable trigger distributed backup mechanisms that counterbalance or nullify the effect of further changes that would have been brought about along P.

In cases of distributed robustness there may be backup mechanisms that mask causal relations by ruling out counterfactual dependence between the cause and effect. We have suggested that such cases can be handled by establishing mediated causal relations that are not grounded directly by counterfactual dependence. This requires a chain of mediating causal relations of the right kind, given by the relevance requirement. In dynamic cases with distributed robustness, how should we think about the relevance requirement and about the truth conditions for the counterfactual dependencies?

Our tentative suggestion is that relevant counterfactuals should be evaluated by looking at systems that are similar to the systems of interest *at the time changes are induced*. When inducing multiple changes at different times, the counterfactual scenarios involve systems that are similar to the system of interest up to the point of the relevant change. Even if it is a variable upstream that we are interested in checking the causal relevance of, we should let the counterfactual target system evolve like the normal system up to the point of changes in downstream variables. In this way we avoid that distributed backups potentially triggered by earlier changes mask the mediated causal relationships we want to reveal. The way to think about truth-conditions for causal counterfactuals about systems exhibiting distributed robustness and non-modular behavior is to compare a normal system with different counterfactual systems subject to the same dynamic evolution as the normal system up to the time of changes of the mediating variables.

This is a tentative definition of causal relevance, in the broad sense, for systems changing their dynamic evolution as a result of perturbations. It is designed to be a special case of the general philosophical analysis of causation that we started out with. There will also be dynamic analogues to restricted and actual causation, by restricting the relevant values to normal values and to actual values respectively. Developing a full-fledged philosophical account along these lines is a task for future work, but we hope to have made a convincing case for the philosophical interest of representing the dynamics of causal systems.

6. Concluding remarks

We have used biological examples of gene-redundancy and distributed robustness to suggest some extensions and revisions of the philosophical understanding of causation. The focus has been on cases of causation where there are no direct variable-on-variable counterfactual dependencies, and we have suggested that changing the resolution of the causal representation is a natural move in such cases. This can be done by increasing or decreasing the resolution of the causal model. Either way you go, causal claims face the tribunal of experience in concert. The relativization to a model puts the focus of causal investigation where it should

be; namely on generating good causal models, rather than establishing singular causal claims in isolation.

References

Cartwright, N., 2001, "Modularity: It Can, and Generally Does, Fail", in: M.C. Galavotti, P. Suppes, and D. Constantini (Eds.), *Stochastic Causality. CSLI Lecture Notes*. Stanford, CA: CSLI Publications, pp. 65-84.

Edwards, J. S. and Palsson B. O., 2000a, "Robustness Analysis of the *Escherichia coli* Metabolic Network", in: *Biotech Progress* 16, pp. 927-939.

Edwards, J. S. and Palsson B. O., 2000b, "The *Escherichia coli* MG1655 *in silico* Metabolic Genotype: Its Definition, Characteristics and Capabilities", in: *Proceedings of the National Academy of Science of the United States of America* 97, pp. 5528-5533.

Hanada, K., Sawada Y., Kuromori T., Klausnitzer R., Saito K., Toyoda T., Shinozaki K., Li W.-H., and Hirai M. Y., 2011, "Functional Compensation of Primary and Secondary Metabolites by Duplicate Genes in Arabidopsis thaliana", in: *Molecular Biology and Evolution* 28, pp. 377-382.

Hausman, D. and Woodward J., 1999, "Independence, Invariance and the Causal Markov Condition", in: *British Journal of the Philosophy of Science* 50, pp. 521-583.

Kitano, H., 2004, "Biological Robustness", in: *Nature Reviews Genetics* 5, pp. 826-837.

Kuznicki K. A., Smith P.A., Leung-Chiu W. M. A., Estevez A. O., Scott H. C. and Bennett K. L., 2000, "Combinatorial RNA Interference Indicates GLH-4 Can Compensate for GLH-1; These Two P Granule Components are Critical for Fertility in *C.* elegans", in: *Development* 127, pp. 2907-2916.

Lewis, D., 1973, "Causation", in: *Journal of Philosophy* 70, pp. 556-567.

Lewis, D., 2004, "Causation as Influence", in: J. Collins, N. Hall and L.A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge (Mass.): The MIT Press, pp. 75-117.

Menzies, P., 2004, "Difference Making in Context", in: J. Collins, N. Hall and L.A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge (Mass.): The MIT Press, pp. 139-180.

Mitchell, S., 2009, *Unsimple Truths. Science, Complexity and Policy*. Chicago: The University of Chicago Press.

Paul, L. A., 2004, "Aspect Causation", in: J. Collins, N. Hall and L.A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge (Mass.): The MIT Press, pp. 205-224.

Shastry, B. S., 1994, "More to Learn from Gene Knockouts", in: *Molecular and Cellular Biochemistry* 136, pp. 171-182.

Strand, A. and Oftedal G., 2009, "Functional Stability and Systems Level Causation", in: *Philosophy of Science* 76, pp. 809-820.

Wagner, A., 2005, "Distributed Robustness versus Redundancy as Causes of Mutational Robustness", in: *BioEssays* 27, pp. 176-188

Woodward, J., 2003, *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J., 2010, "Causation in Biology", in: *Biology and Philosophy* 25, pp. 287-318.

Xie, J., Awad K. S., Guo Q., 2005, "RNAi Knockdown of Par-4 Inhibits Neurosynaptic Degeneration in ALS-linked Mice", in: *Journal of Neurochemistry* 92, pp. 59-71.

Anders Strand
Department of Philosophy, Classics, History of Art and Ideas
University of Oslo
Box 1020 Blindern
0315, Oslo
Norway
anders.strand@ifikk.uio.no

Gry Oftedal
Department of Philosophy, Classics, History of Art and Ideas
University of Oslo
Box 1020 Blindern
0315, Oslo
Norway
gry.oftedal@ifikk.uio.no

Melinda Bonnie Fagan

Experimenting Communities in Stem Cell Biology: Exemplars and Interdisciplinarity

Abstract

This essay uses three case studies to illustrate the importance of experimenting communities in stem cell biology. An experimenting community is a collection of scientific groups that together produce knowledge using experimental methods. Three such methods, each an exemplar for stem cell biology, reveal the structure and significance of experimenting communities in stem cell research: the spleen colony assay, embryonic stem cell lines, and systems models. Together, these case studies show that (1) stem cell research progresses via multiple, diverse models and comparisons among them; (2) the spleen colony assay and embryonic stem cell lines have a special status in this field, as hubs of experimental networks; (3) hierarchical cell lineage models are a unifying framework for stem cell biology today; and (4) another general model of development, Waddington's landscape, can help merge stem cell and systems biology into a new, expanded, experimenting community.

1. Introduction

This essay uses three case studies to illustrate the importance of experimenting communities in stem cell biology. An experimenting community is a collection of scientific groups that together produce knowledge using experimental methods. In stem cell biology, knowledge takes the form of robust models and detailed mechanistic explanations. The cases discussed below show that these are not individual accomplishments, but emerge from a network of experimental systems, organized into a community with shared exemplars and standards. To understand stem cell biology as a science, we need to attend not only to individuals and research teams, but also the wider communities to which they belong. Because these communities do not map smoothly onto traditional disciplinary divisions, issues of interdisciplinarity arise as well. Indeed, stem cell biology is a particularly rich site for exploring interdisciplinary issues, as its central research program undercuts traditional dualisms of science/medicine and science/technology.

Each case focuses on an exemplary method for stem cell research: the spleen colony assay, embryonic stem cell lines, and systems models. The spleen assay provided the first direct evidence of stem cells in mammals, in 1961. Embryonic

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_16,

stem cell lines were created 20 years later, but assumed their current exemplary status more recently, in 1998. Systems models for stem cells are still in very early stages, part of the rising field of systems biology. So this essay traces the history of stem cell biology through its exemplars. Before delving into the cases, some background is necessary. I first set out a minimal consensus definition of 'stem cell,' and introduce the core concepts and methods of stem cell biology today. These general considerations motivate the more detailed case studies that follow. Together, the case studies illustrate four points: (1) robust knowledge about stem cells emerges from a network of comparisons among diverse models and experimental methods. (2) The spleen colony assay and embryonic stem cell lines occupy central positions within this experimental network. (3) Cell lineage hierarchy is unifying pattern across the network, which coordinates diverse mechanisms from different experimental systems. And, (4) another general model of development, Waddington's landscape, can help merge stem cell and systems biology into a new, expanded, experimenting community.

2. Identifying stem cells

Stem cells today are defined as undifferentiated cells that self-renew and give rise to differentiated cells. This 'consensus concept' involves two reproductive processes: self-renewal and differentiation. Self-renewal is production of more cells like the parent with respect to some set of traits of interest, for a given duration. Differentiation is production of cells unlike the parent and one another. So differentiation has two dimensions: diversity of cells at a time, and change over time. These comparisons are also relative to some set of traits. A simple, minimal way to represent the stem cell concept is as a cell hierarchy organized by these two reproductive relations (Figure 1). In this minimal model, a stem cell is defined by position in a cell hierarchy: the unique stem of a cell lineage. Within a lineage, and relative to a set of traits and a temporal duration of interest, a stem cell has maximal self-renewal and differentiation potential. So the stem cell concept, minimally-modeled in this way, is relational and relative.

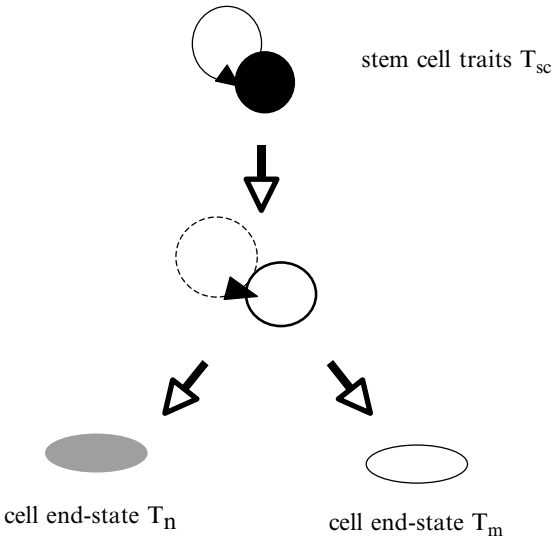


Figure 1. Minimal stem cell model.

Experimental methods for identifying stem cells share a basic pattern: remove cells from an organismal source, place them in a context in which their traits can be measured, then move cells to another environmental context to measure stem cell capacities (Figure 2). Any particular method that conforms to this pattern specifies parameters that the minimal model leaves open.

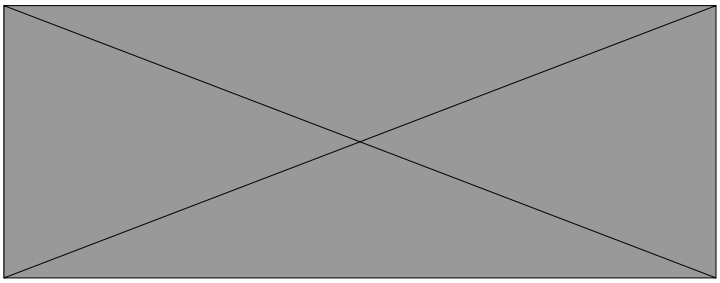


Figure 2. Basic structure of experiments that aim to identify and characterize stem cells.

Organismal sources differ in species, developmental stage, and site from which cells are extracted. Cells from a given source differ in their physical, molecular and morphological traits. Differentiation potential and self-renewal are

measured, in a new context, and correlated with the first set of traits. The result of such an experiment is a mapping between traits of an organismal source, extracted cells, and differentiated cells to which the former give rise under controlled conditions. Different kinds of stem cell differ with respect to organismal source, set of traits, self-renewal duration, or scope of differentiation potential.

Generalizations about stem cell capacities are relative to experimental methods. This is because self-renewal and differentiation potential are not measured directly, but inferred from observations of offspring: stem cells are revealed by their descendants. Because cells reproduce by division, descendants and ancestors cannot co-exist. It follows that a single cell cannot be experimentally shown to be capable of *both* self-renewal and differentiation. To establish its differentiation potential, a candidate stem cell must be placed in an environment conducive to differentiation, and its progeny observed. To establish its self-renewal capacity a cell must be maintained in an environment that is *not* conducive to differentiation, and its progeny observed. It is not possible to perform both experiments on a single cell. So assignment of stem cell capacities to single cells is necessarily uncertain. Nor can differentiation potential of a single cell be experimentally demonstrated. Experiments can reveal what happens when a cell is placed in a particular environment, but not what would have happened if that cell had been placed in a different environment. Given pure populations of identical stem cells, experiments can show what happens to a stem cell of given type in a range of environments. But we still have no basis for generalizing beyond the range of environments used in experiments. And the assumption of pure populations of identical stem cells is, at best, provisional. We do not have such populations in hand, in advance of experiments that identify stem cells, which are the very experiments at issue.

Therefore, stem cells can be identified and characterized only relative to an experimental method, which includes a specific organismal source, a set of traits measured, and manipulations of cells' environments. In consequence, experimental methods, rather than lawlike generalizations or overarching principles, are the main epistemological focus for stem cell biology. The primary inferential task in this field is not generalization from one exemplar, but coordination of diverse model systems. The next sections illustrate these abstract points in concrete detail.

3. Case 1: Spleen colony assay[1]

The spleen colony assay provided the first experimental demonstration of stem cells in mammals (Figure 3). The method conforms to the general pattern outlined previously. The organismal source is adult mouse bone marrow. Bone marrow cells, which give rise to the immune system, are highly sensitive to radiation. But irradiated mice can be 'rescued' with a transplant of bone marrow cells, which

1 This section builds on Fagan 2007, 2010, 2011.

effectively transplants the entire blood and immune system. The first manipulation in this exemplary method is to collect cells from mouse bone marrow; the first measurement is to count them. The second manipulation is to inject a known number of cells into lethally-irradiated mice. After about a week, spleens are removed from surviving mice. As a byproduct of 'radiation rescue,' their spleens display lumpy nodules of $>10^6$ cells: spleen colonies. The second measurement is to count the number of spleen colonies. Experimental results are pairs of datapoints, correlating number of cells injected to number of spleen colonies observed.

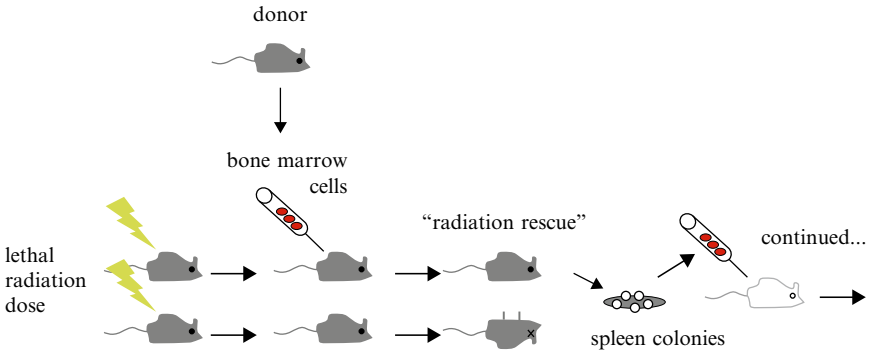


Figure 3. Sketch of the spleen colony assay

The spleen colony assay emerged from the interface of several experimenting communities: radiation biology, cancer research, and molecular biology (Brown et al. 2006, Fagan 2007, Kraft 2009, Fagan 2010). In the early 1960s, James Till and Ernest McCulloch, working at the Ontario Cancer Center in Toronto, made the crucial connection between radiation rescue in mice, and contemporary findings of tissue culture and molecular biology. Their key insight was that an irradiated mouse is analogous to a tissue culture environment, and splenic nodules to bacterial colonies of bacteria, each derived from a single cell:

> the irradiated mouse may be considered as providing the receptacle, the medium, and control of temperature, pH, and humidity required for the cultivation of marrow cells… [Spleen colonies offer] a direct method of assay for these [normal mouse bone marrow] cells with a single-cell technique (Till and McCulloch 1961, pp. 220, 213).

Unlike bacteria colonies, however, spleen nodules contained several different types of blood cell, as well as cells that could give rise to more spleen colonies. So the assay demonstrated that mouse bone marrow contained cells capable of extensive proliferation, differentiation into multiple blood cell types, and self-renewal. Based on these features, spleen colony-forming cells were identified as blood stem cells in mice (mHSC).

Though the method was innovated by a single group, a new experimenting community soon developed around the assay, with centers in Toronto, Melbourne, Rijswijk, Manchester, and the Eastern US. After 1961 these centers 'differentiated,' inventing new assays; variations on the spleen colony theme. The aim of this worldwide effort was to characterize and control colony-forming HSC, and to extend these methods to humans. Over two decades, this research effort produced the blood stem cell hierarchy, a compendium of knowledge about blood cell development and an exemplary model for adult stem cell research. But the process was not a straightforward collaboration involving a simple division of labor ('one model, one community'). The concrete models and methods at different HSC centers diverged widely. Some focused on specific blood cell lineages, others on the biochemistry of regulatory factors in blood cell development, still others on the concept of a stem cell niche. Intersection and merging of different experimenting communities played a crucial role in constructing our present model of HSC development.

One key 'merger' was between the community that formed around the spleen colony assay, and cellular immunology. In the former, efforts to isolate colony-forming cells concentrated in the Netherlands, at the Radiobiological Institute in Rijswijk. The Rijswijk group coalesced around a clinical goal: improving efficiency of bone marrow transplantation, then and now a treatment for leukemia and other pathologies. Their strategy was to find a combination of "well-defined" cell properties that correlated with blood stem cell capacities, allowing researchers (and eventually, clinicians) to select a population of all and only HSC. Rijswijk researchers scoured the published literature for ways of labeling blood cells that corresponded to known biophysical or molecular characters, sorted bone marrow cells into 'subsets' accordingly, tested each subset for stem cell capacities using the spleen colony assay, and finally, characterized subsets microscopically (Figure 4). The result was an array of well-characterized cell populations, each with a quantifiable degree of HSC function.

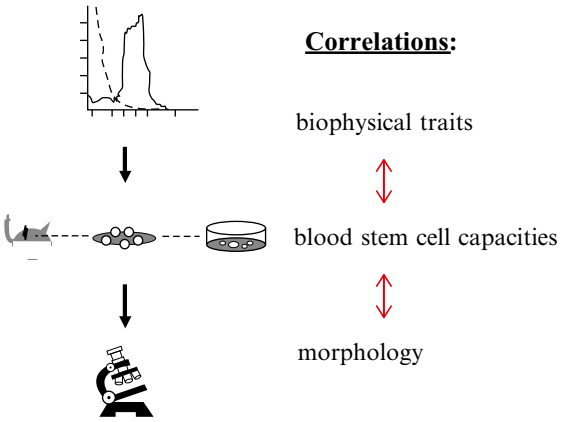


Figure 4. Sketch of Rijswijk method for identifying mHSC (1961–1988) and results.

The Rijswijk group kept pace with the wider community by incorporating newly-defined traits into an ever-lengthening protocol, but otherwise their experiments were a 'closed system.' There were at least two reasons for this. First, their method was designed to yield results that could be rapidly translated to clinical contexts. For rapid translation, well-defined cell properties, clear criteria for sorting cell subsets, and use of technologies belonging to a "biomedical platform" were crucial.[2] So the basics of the Rijswijk method could not be altered without compromising the group's overall goal. Second, this goal was not widely shared in the experimenting community that developed around the spleen colony assay. That community was highly fragmented, such that groups comprising it shared little apart from a historical tie to the spleen colony assay. It made sense for the Rijswijk group to treat the rest of the community as a resource for identifying new cell traits for use in their method.

However, as the Rijswijk researchers made steady progress toward characterizing mHSC on their own terms, in the wider community key background assumptions were changing – including the definition of HSC.[3] Over the 1970s and early 1980s, identification of HSC with spleen colony-forming cells gave way to a "three-tiered model" of HSC, lineage-committed progenitor cells, and mature blood cells. Gradually, the Rijswijk group's standard diverged from that accepted in the wider HSC community. This was, in part, due to their distinctive goal, which oriented them more toward engagement with clinical practice than with other groups seeking to understand HSC. So when, in 1984, the Rijswijk group described a method for isolating all and only mHSC, there was little response. But a nearly identical announcement by a Stanford group, four years later, became

2 For more on biomedical platforms, see (Keating and Cambrosio 2003).

3 Fagan 2011 describes these changes in more detail.

a landmark of blood stem cell research, and the basis for HSC models today.[4] Understanding this situation requires attention to both the fine-grained details of experimental methods, the overall goals of each group, and their modes of participation in a wider community.

The two groups differed in experimental standards, goals, and community organization. To estimate HSC enrichment in their results, the Stanford group used long-term survival after radiation, while the Rijswijk group used number of spleen colonies. Unlike the Rijswijk institute, the Stanford group did not have direct clinical aspirations, but focused instead on contributing to the emerging field of cellular immunology. So the two laboratories initially belonged to different experimenting communities: hematologists (blood cell experts) and immunologists (focused on cells of the immune system specifically). The Stanford group's method was successful not for the cells it isolated, but for bringing about a fruitful merger of these experimenting communities. It accomplished this by prioritizing collaboration and focusing experiments on immune cell development.

The Stanford laboratory, headed by Irv Weissman, was a hub for studying immune cell development. Lab members were encouraged to collaborate, both inside and outside the lab. Through this network of collaborations, lab members helped to chart developmental pathways for immune cells. Progress was made by merging independently-constructed models of cell developmental pathways, as common ancestors were identified across experimental systems. The 1988 report was based on exactly this strategy: knitting together previously distinct research projects into a single coordinated search for HSC.[5] After publication, essentially the same process occurred at the level of experimenting communities. The 1988 experiment characterized HSC as a point of convergence among diverse projects, representing its relations to other blood cell types in a branching tree of cell development. This model of blood cell development corresponded well to the three-level hierarchy that had come to be widely-accepted in the hematology community. The two communities merged to form a new, expanded HSC community, with ties to both immunology and hematology.

By prioritizing collaboration with other cellular immunologists over direct clinical applications, the Stanford group established a robust and inclusive framework for research on cell development. The HSC model and method were quickly extrapolated to other experimental systems: from mouse to human HSC; from blood to brain, gut, skin, muscle, liver, pancreas, the enteric nervous system; from normal to cancerous development, including leukemia, colon cancer, breast cancer, and prostate cancer. For each new system, HSC experiments and results served as an exemplar and a basis for comparison. Insights in other systems in turn 'fed

4 Visser et al. 1984, Spangrude et al. 1988, respectively. It is important to note that there is no widely agreed-upon procedure for isolating all and only HSC. HSC are method-relative: unless you know how the cells were isolated and sorted, you don't know what they are.

5 For more detail, see (Fagan 2007 and 2010).

back' to further refine the HSC model, elaborating the simple three-level hierarchy to an intricately-tiered structure of lineages. In this way, the epistemic community of adult stem cell research grew as a reticulated network of experimental systems.

4. Case 2: Embryonic stem cell lines

The same pattern is seen in the other branch of stem cell research, focused on embryonic cells. Embryonic stem cell (ESC) lines play a role analogous to HSC. The method of producing human ESC conforms to the basic experimental pattern set out in Figure 2. The organismal source is the inner cell mass of a 5-day human blastocyst. The first manipulation is to remove part of this inner layer and place it in artificial cell culture containing nutrients and factors that prevent differentiation. Cells that rapidly divide under these conditions form colonies, which are selected and put into new cultures. This "passaging," repeated every few weeks, maintains a continuously-growing lineage of undifferentiated human cells: a human embryonic cell line. The first measurement is of cell traits in these cultures: morphology, surface molecules, and DNA/RNA/protein sequences. Next, samples of the growing cell line are transferred to an environment that encourages differentiation. By varying cell culture conditions, differentiation can be biased toward a particular cell type: neurons, cardiac muscle, blood, *etc*. The second measurement is of these differentiated cells' traits.

The ESC method was invented two decades after the spleen colony assay, and shares its roots in cancer research. However, its exemplary status dates only from 1998, when the method was first successfully applied to human cells. This was accomplished by an international research team led by James Thomson and colleagues at the University of Wisconsin. Though technically challenging, the method that yielded the 1998 breakthrough was strikingly *unoriginal*. It was adopted, with very little alteration, from a procedure used in mice nearly two decades earlier, which was in turn, a modification of methods pioneered in the 1950s. All these techniques produced self-renewing, pluripotent stem cell lines, which differ mainly in organismal source. The earliest were derived from a form of testicular cancer, teratocarcinoma, which produces tumors composed of diverse cell types. This method was streamlined in the 1970s to produce embryonal carcinoma cell lines (EC), then applied to normal mouse embryos to yield mESC lines (Evans and Kaufman 1981, Martin 1981). So embryonic stem cell lines derived from mice have existed since 1981. But they did not become an exemplary focus for a new scientific field. Instead, mouse ESC were primarily used as tools for genetically manipulating the mammalian germline – part of the technology for making 'knockout' mice.

ESC lines were established as an exemplar when the method was extended to humans, after preliminary work on other primate species (Thomson et al. 1998).

The method itself was essentially the same, apart from details of measured cell traits. What was original, however, was explicit articulation of a therapeutic goal:

> The standardized production of large, purified populations of euploid human cells such as cardiomyocytes and neurons will provide a potentially limitless source of cells *for drug discovery and transplantation therapies* … Progress in basic developmental biology is now extremely rapid; human ES cells will link this progress even more closely to the *prevention and treatment of human disease* (*ibid*., pp. 1146-7; italics mine).

Like their mouse counterparts, hESC were conceived as tools. However, their purpose was to help realize therapeutic goals, as well as to reveal genetic pathways of early development. The explicitly clinical aim unified and galvanized a new biomedical research community, which adopted the ESC method as a standard: create a cultured cell line with unlimited self-renewal and potential to differentiate into all cell types of the adult organism (pluripotency). The embryonic branch of stem cell research resulted from this merger of mammalian genetics and development with clinical research. In this sense, the field is inherently 'translational': aimed at establishing efficient connections between bench and bedside. The best methods for such a translational field are still a topic of spirited debate, complicated by the multiple ethical dimensions involved.

Since 1998, pursuit of this clinical goal has yielded a second constellation of model systems: different pluripotent stem cell lines, produced by often quite subtle variations in the timing or location of extraction from an organismal source (ESC, EpiSC, EC, GSC, and, most recently iPSC; see below). Just as for tissue-specific counterparts, robust results about pluripotent stem cells emerge by comparing results of different experiments. However, human ESC, like mouse HSC, play a distinctive, privileged role within this network of comparisons. There are several reasons for this. First, of the human cell lines, ESC consistently show the widest differentiation potential, lowest tumor formation, need for minimal manipulation, and highest reproductive rate. All these features are desirable in light of the therapeutic goal that motivates and unifies embryonic stem cell research. Their significance is epistemic as well as practical, as better tools will enable us to gain clinically relevant knowledge more quickly. Human ESC lines exhibit the desired stem cell capacities to the greatest extent and in the most accessible way; in this sense, they are 'exemplary models' of early human development. They are also *entrenched*. Most of the past decade or so of research on pluripotent stem cells is articulated in relation to hESC capacities. Since our knowledge about stem cells emerges from comparisons among different model systems, to excise hESC would eliminate the epistemic core of pluripotency research, impoverishing the entire field.

Opponents of research using human embryos have repeatedly argued that tissue-specific stem cells, such as those in adult bone marrow or umbilical cord blood, are just as clinically promising as hESC, making the latter unnecessary.

This argument rests on a misunderstanding of the dynamics of experimenting communities. Any scientific field makes progress by building on past success. In both branches of stem cell biology, past success involves coordinating diverse model systems, so their results can be compared and integrated. This is accomplished by merging different experimenting communities. If the past is any guide, then adult and embryonic branches should merge into a community that can produce a more comprehensive account of cell development. As a step in this direction, it is interesting to note that the goals associated with the founding exemplars of two branches of stem cell research are complementary. Adult stem cell research acquired its present configuration by placing clinical goals in the background, a long-term eventual aim rather than a consideration that structured experiments in the short-term. In contrast, embryonic stem cell research acquired its present configuration by shifting focus from genetic dissection of mammalian development to therapeutic applications and drug development. The recent shared emphasis on translational research, together with experimental methods involving both kinds of stem cell, indicates the two branches are now poised to merge into a single experimenting community. It is a mistake to conceive of them as competing alternatives.[6]

5. Case 3: Systems models[7]

My third case also concerns the future of stem cell biology. Systems models are not (yet) an exemplar, but part of the developing interface of stem cell and systems biology. Since both fields aim to explain cell development, scientists are increasingly concerned with their relation. The still-murky intersection of stem cell and systems biology is a site at which philosophers of science can make a valuable contribution. My suggestion is that a simple interfield model, Waddington's epigenetic landscape, can facilitate a productive merger of stem cell and systems biology. This simple model is a two-dimensional diagram representing pathways of organismal development as a three-dimensional strucure (Waddington 1957; Figure 5). The axis projecting outward represents time. The horizontal axis represents phenotype ordered by similarity (the measure of which is left unspecified). The vertical axis represents order of development, correlated with time such that the developmental surface is tilted. In this model, development is a force that operates like gravity on organisms and their parts. But its operation is structured into branching tracks, with a single starting point leading to multiple end-states. The landscape, as a whole, represents developmental options, progressively restricted over time. If the developing entity is interpreted as a cell, the landscape model

6 As argued in the August 2010 injunction on federally-funded hESC research in the US, for example.

7 For more detail see (Fagan 2012).

visualizes the stem cell concept: a single undifferentiated starting point, with the potential to develop along a variety of pathways, gradually restricted as development proceeds, ending with stable, mature cell types.

This model of development is particularly salient for 'reprogramming' experiments, which manipulate cells' developmental potential. In these experiments, differentiated cells are extracted from an organism and cultured with a few specific genes. After several weeks, about 0.5% of these cells resemble ESC, morphologically and functionally: they can self-renew, and give rise to many different cell types. These "reprogrammed" cells are termed iPSC, induced pluripotent stem cells. This experiment was first successfully performed in 2006, by a team from Kyoto University led by Shinya Yamanaka (Takahashi and Yamanaka 2006). Over the past five years, their method has been replicated and refined by thousands of studies, with variations in many experimental parameters.[8] 'Reprogrammers' use Waddington's landscape not to represent individual experiments, but when 'pulling back' to consider the collective community effort. In these contexts, the model plays several distinct roles. One is to visualize shared background assumptions, such as the stem cell concept. Another is to summarize the results of many experiments, so as to suggest explanations for these generalizations. In this way, the landscape model helps to coordinate different reprogramming experiments. Finally, reprogrammers use the landscape to correlate developmental potential with molecular cell state (Figure 6). This usage offers a potentially fruitful link with systems biology.

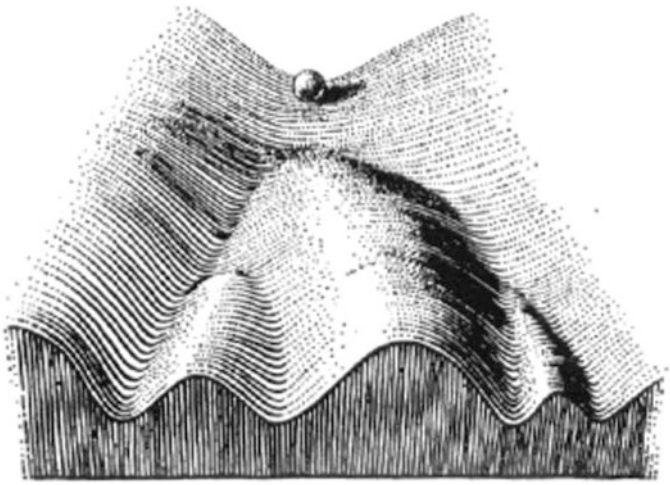


Figure 5. Waddington's epigenetic landscape (from Waddington 1957, p. 29).

8 See (Maherali and Hochedlinger 2008) for a comprehensive, though now dated, review. Note that this method conforms to the basic pattern of Figure 2, and is explicitly modeled on the method for producing ESC lines (Case 2).

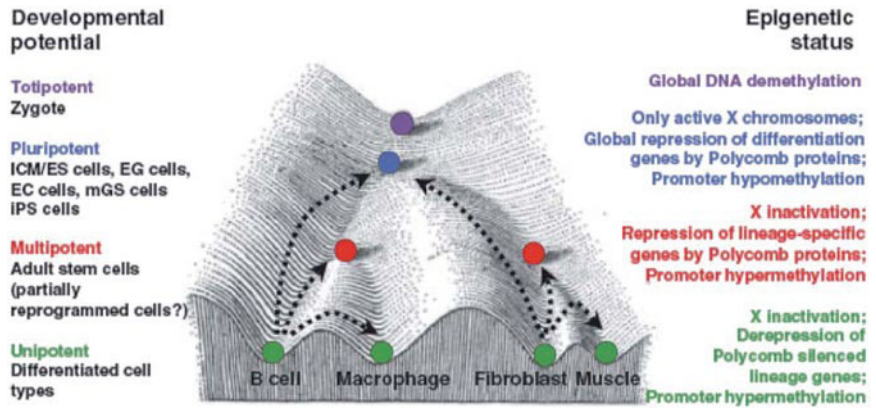


Figure 6. Cooption of Waddington's landscape to represent cell reprogramming experiments (from Hochedlinger and Plath 2009, p. 510, reproduced with permission of *Development*)

A central tenet of systems biology is that underlying molecular networks control cell behavior. Cell development is thought to be controlled by 'transcription networks' that determine which DNA sequences are transcribed in a cell at a given time. These networks can, in principle, be mathematically modeled as consisting of a finite set of molecular elements (genes, RNA, protein and small molecules), each characterized by a value of the state variable at time *t*. Cell behavior (including development) is understood as the result of changes in the values of these variables. Systems models represent the relations among variables comprising the cell state in a formal framework, usually a system of ordinary differential equations that describe how the state variable changes over time. Solutions to the system of equations define features of the 'landscape' in state space. One kind of solution that can often be found are 'steady-states', at which there is no change in values of variables. Another are 'local solutions' given initial conditions; i.e., values for the state variable of each molecular species at a given time *t* (corresponding to a point in state space). Each local solution describes exactly how the system as modeled will change over the next small increment of time. They are represented as vectors, tiny paths in state space.

The arrangement of vectors and steady-states in state space determines landscape topography. A stable steady-state is one toward which vectors converge – a 'stable attractor' – while an unstable steady-state is one from which vectors radiate away.[9] The landscape is produced by adding this dimension of 'stability' to the state space. For a two-dimensional state space, this is visualized as 'elevation' of the landscape surface. A systems model of this kind entails many predictions about how the system as a whole will respond to manipulations that change the

9 See (Klipp et al. 2009) and (Szallasi et al. 2010).

state variable of one or more elements. Given the correlation of cell state and developmental potential, and a 'developmental order' identifying the stem state, systems models can derive the developmental landscape of a cell from the bottom-up. Such a derivation yields a rigorous, multi-level explanation of cell development, grounded jointly in experimental manipulation and mathematical modeling. Waddington's landscape offers a 'blueprint' for interdisciplinary collaboration aimed at articulating robust and useful explanations of cell development.

6. Conclusion

I have used three case studies to demonstrate the importance of experimenting communities in stem cell biology. Each focuses on a different exemplar: the spleen colony assay in mice, human embryonic stem cell lines, and mathematical systems models of single-cell gene expression. Together, these cases yield several key results. First, because stem cell capacities can be attributed to cells only relative to experimental methods, the field progresses by multiplying models and constructing a network of comparisons among them. Within this experimental network, the spleen colony assay and embryonic stem cell lines have a special status as foundations of the field in its current configuration. As for models, the cell lineage hierarchy is a stable and robust pattern. It operates as a framework to coordinate and link diverse mechanisms of cell development worked out in different experimental systems. But, going forward, a third dimension should be added – bringing the cell lineage hierarchy into alignment with Waddington's landscape. In this way, stem cell and systems biology can merge to form a new experimenting community, combining concrete experiments and mathematical modeling to explain cell development.

Acknowledgements: Early stages of this research were funded by the National Science Foundation (grant number 0620993), and later stages by a Mosle Research Fellowship (2009–2011), a Rice Humanities Research Center Collaborative Fellowship (2009–2010), and a Rice Faculty Innovation Grant (2010-2012). Many of these ideas were honed in conversation with scientists and philosophers; special thanks to Leo Aguila, Colin Allen, Mike Clarke, Jordi Cat, Hasok Chang, Tom Gieryn, Richard Grandy, Len and Lee Herzenberg, Oleg Igoshin, Libuse Jerabek, Motonari Kondo, Hannah Landecker, Elisabeth Lloyd, Helen Longino, Sean Morrison, Christa Müller-Sieberg, Jutta Schickore, Fred Schmitt, Paul Simmons, Jerry Spangrude, Amy Wagers, and Irv Weissman. The project has also benefited from the work of excellent undergraduate research assistants: Tracey Isidro, Dandan Liu, and Casey O'Grady. An earlier version of this paper was presented at the Interdisciplinarity and Systems Biology Workshop in Aarhus, Denmark (August

18, 2011); many thanks to Hanne Andersen and other participants for helpful comments and criticism.

References

Brown, N., Kraft, A., and Martin, P., 2006, "The Promissory Pasts of Blood Stem Cells", in: *BioSocieties* 1, pp. 329-348.

Evans, M., and Kaufman, M., 1981, "Establishment in Culture of Pluripotential Cells from Mouse Embryos", in: *Nature* 292, pp. 154-156.

Fagan, M., 2007, "The Search for the Hematopoietic Stem Cell: Social Interaction and Epistemic Success in Immunology", in: *Studies in History and Philosophy of Biological and Biomedical Sciences* 38, pp. 217-237.

Fagan, M., 2010, "Stems and Standards: Social Interaction in the Search for Blood Stem Cells", in: *Journal of the History of Biology* 43, pp. 67-109.

Fagan, M., 2011, "Social Experiments in Stem Cell Biology", in: *Perspectives on Science* 19, pp. 235-262.

Fagan, M., 2012, "Waddington Redux: Models and Explanation in Stem Cell and Systems Biology", in: *Biology and Philosophy* 27, pp. 179-213.

Hochedlinger, K., and Plath, K., 2009, "Epigenetic Reprogramming and Induced Pluripotency", in: *Development* 136, pp. 509-523.

Keating, P., and Cambrosio, A., 2003, *Biomedical Platforms: Realigning the Normal and the Pathological in Late-twentieth-century Medicine.* Cambridge (Mass.): The MIT Press.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach H., and Herwig, R., 2009, *Systems Biology: A Textbook*. Weinheim: Wiley-VCH.

Kraft, A., 2009, "Manhattan Transfer: Lethal Radiation, Bone Marrow Transplantation, and the Birth of Stem Cell Biology, ca. 1942–61", in: *Historical Studies in the Natural Sciences* 39, pp. 171-218.

Maherali, N., and Hochedlinger, K., 2008, "Guidelines and Techniques for the Generation of Induced Pluripotent Stem Cells", in: *Cell Stem Cell* 3: 595-605

Martin, G., 1981, "Isolation of a Pluripotent Cell Line from early Mouse Embryos Cultured in a Medium Conditioned by Teratocarcinoma Stem Cells", in: *Proceedings of the National Academy of the Sciences USA* 78, pp. 7634-7638.

Spangrude, G., Heimfeld, S., and Weissman, I., 1988, "Purification and Characterization of Mouse Hematopoietic Stem Cells", in: *Science* 241, pp. 58-62.

Szallasi, Z., Stelling, J., and Periwal, V. (Eds.), 2010, *Systems Modeling in Cell Biology: From Concepts to Nuts and Bolts*. Cambridge (Mass.): The MIT Press.

Takahashi, K., and Yamanaka, S., 2006, "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors", in: *Cell* 126, pp. 663-676.

Thomson, J., Itskovitz-Eldor, J., Shapiro, S., Waknitz, M., Swiergiel, J., Marshall, V., Jones, J., 1998, "Embryonic Stem Cell Lines Derived from Human Blastocysts", in: *Science* 282, pp. 1145-1147.

Till, J. and McCulloch, E., 1961, "A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells", in: *Radiation Research* 14, pp. 213-222.

Visser, J., Bauman, G., Mulder, A. Eliason, J., and de Leeuw, A., 1984, "Isolation of Murine Pluripotent Hemopoietic Stem Cells", in: *Journal of Experimental Medicine* 59, pp. 1576-1590.

Waddington, C. H., 1957, *The Strategy of the Genes*. London: Taylor & Francis.

Department of Philosophy
Rice University
P.O. Box 1892
Houston, Texas 77251-1892
USA
mbf2@rice.edu

William Bechtel

From Molecules to Networks: Adoption of Systems Approaches in Circadian Rhythm Research

Abstract

In the 1990s circadian rhythm researchers made enormous progress in identifying the components and operations within the responsible mechanism in various species using the tools of molecular biology. In the past decade it has proven essential to supplement these with the tools of systems biology both to identify additional components but especially to understand how the mechanism can generate circadian phenomena. This has proven especially important since research has shown that individual neurons in the mammalian mechanism are highly variable and that the way they are organized in networks is crucial to generating regular circadian behavior.

1. Introduction

From its roots in the study of circadian rhythms observed in physiology and behavior, circadian rhythm research rapidly adopted and energetically pursued a molecular biological approach in the last decades of the 20th century. This research has been highly productive in revealing many of the components of the circadian mechanisms in each of the major model systems: cyanobacteria, fungi, plants, and various animals (especially fruit flies and mice). But success in decomposing the mechanisms has also generated challenges in recomposing them, a crucial step in understanding how they work. Although in some fields it is possible for researchers to literally recompose mechanisms (e.g., by reconstituting a chemical reaction in vitro), in other fields researchers must do so more indirectly, either by imagining the interactions of the components performing their various operations or by constructing computational models that demonstrate how the hypothesized set of components would interact if they operated in the manner characterized. Imagination suffices when mechanisms are relatively simple, involving components performing linear operations and organized sequentially. But when the parts identified operate non-linearly and are organized non-sequentially, such an approach

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_17,

fails. The alternative, increasingly being pursued in circadian rhythm research, is to turn to computational modeling and dynamical systems analysis.[1]

A further challenge stems from the fact that underlying the strategy of decomposing mechanisms is the assumption that the mechanism itself and each of its components operate largely in isolation from other mechanisms or components so that the whole system exhibits what Herbert Simon referred to as *near decomposability*.[2] Assuming near decomposability is a heuristic, and a characteristic of heuristics is that they can fail. Increasingly biologists are learning that the mechanisms they study are less decomposable then they thought, and circadian mechanisms are no exceptions. The challenge is to relax the decomposability assumption and incorporate the influences from other components that alter the behavior of the components into one's account without losing the ability to explain the operation of the mechanism in terms of its components. Once again, this is leading circadian researchers to turn to computational modeling, which has the resources to characterize multiple interactions affecting individual components while they operate within a mechanism.

My focus in this paper will be on the steps in recomposing circadian mechanisms in the last decade that has led to a focus on networks at various levels of organization, including ones at which clock mechanisms interact with other biological mechanisms. This has resulted in an increased focus on networks as opposed to individual components and on the employment of tools from systems biology to understanding the responsible mechanisms. Before examining these developments, though, I will set the stage by introducing circadian rhythms research and briefly describing the results of the more traditional mechanist project of decomposing circadian mechanisms.

---

1 Mechanisms and mechanistic explanation has been the focus of considerable discussion in recent philosophy of science. See, for example, William Bechtel and Robert C. Richardson, *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research.* Cambridge (Mass.): The MIT Press. 1993 edition published by Princeton University Press 1993/2010; Peter Machamer, Lindley Darden, and Carl F. Craver, "Thinking About Mechanisms", in: *Philosophy of Science* 67, 2000, pp. 1-25. In recent papers I have distinguished basic mechanistic explanation, which focuses on recomposing mechanisms through mental simulation, and dynamic mechanistic explanation, which appeals to computational models and dynamical systems theory to recompose mechanisms and explain how they function. See William Bechtel, "Mechanism and Biological Explanation", in: *Philosophy of Science* 78, 4, 2011, pp. 533-557; William Bechtel and Adele Abrahamsen, "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science", in: *Studies in History and Philosophy of Science Part A* 41, 3, 2010, pp. 321-333.

2 Herbert A. Simon, "The Architecture of Complexity: Hierarchic Systems", in: *Proceedings of the American Philosophical Society* 106, 1962, pp. 467-482.

2. From circadian rhythms to clock mechanisms

Circadian rhythms involve endogenously generated oscillations of approximately 24 hours (hence the term *circadian* from *circa* [about] + *dies* [day]) that affect a wide variety of physiological processes and behaviors. For example, human body temperature is lower during the night and raises during the day, varying by nearly a degree Celsius. These rhythms are entrainable to the local day-night cycle; when entrainment cues such as daylight are lacking, they free-run and thereby reveal that their period is not exactly 24 hours. This was one of the crucial features of circadian rhythms that convinced the pioneer circadian researchers in the middle of the 20th century that these rhythms were endogenously maintained and not responses to external cues. The evidence presented at the 1960 Symposium on Biological Clocks at Cold Springs Harbor largely settled the question of endogenous origin of circadian rhythms.[3] While the mechanistic metaphor of a clock was widely embraced by many researchers and employed in the title of the 1960 symposium, the tools for actually investigating the clock mechanism were indirect, relying on such approaches as varying the period of the light-dark cycle or restricting light exposure to pulses at different parts of the cycle to see how they affected the mechanism.

In the two decades after 1960 a variety of researchers identified the locus and began decomposing the hypothesized clock. Although in single-cell organisms and in plants researchers assumed the mechanism was found in each cell, animal researchers assumed that the clock was localized within the brain. Richter discovered that lesions to the hypothalamus disrupted circadian behavior and concluded that circadian rhythms were generated "somewhere in the hypothalamus."[4] In 1972 two research groups further narrowed the locus to the suprachiasmatic nucleus (SCN), a bilateral nucleus located just above the optic chiasm that in the mouse consists of approximately 20,000 neurons. It was the target of projections from the retina, allowing for entrainment by light,[5] and lesions to it rendered animals arrhythmic.[6] Inouye and Kawamura showed, using multi-electrode record-

3 This conference in many respects marks the founding of circadian rhythm research as a distinct research field. The papers and some of the discussion were published in *Cold Spring Harbor Symposia on Quantitative Biology* 25, 1960.

4 Curt P. Richter, *Biological Clocks in Medicine and Psychiatry*. Springfield, IL: Charles C. Thomas 1965.

5 Robert Y. Moore and Nicholas J. Lenn, "A Retinohypothalamic Projection in the Rat", in: *The Journal of Comparative Neurology* 146, 1, 1972, pp. 1-14.

6 Friedrich K. Stephan and Irving Zucker, "Circadian Rhythms in Drinking Behavior and Locomotor Activity of Rats Are Eliminated by Hypothalamic Lesions", in: *Proceedings of the National Academy of Sciences (USA)* 69, 1972, pp. 1583-1586; Robert Y. Moore and Victor B. Eichler, "Loss of a Circadian Adrenal Corticosterone Rhythm Following Suprachiasmatic Lesions in the Rat", in: *Brain Research* 42, 1972, pp. 201-206.

ing, that isolated SCN tissue remained rhythmic.[7] The case for this locus was made more compelling when in 1990 Ralph, Foster, Davis, and Menaker demonstrated that transplanting the SCN from a mutant hamster with a shortened rhythm into ventricles of a SCN-lesioned host restored rhythms in the recipient that corresponded to those of the donor.[8]

To explain how a localized mechanism could function as a clock, researchers needed to decompose it to identify its component parts and the operations they performed. This research proceeded independently using fruit flies during the same period as mammalian researchers were localizing the mammalian clock in the SCN. Since investigators beginning with Darwin viewed circadian rhythms as inherited, a natural strategy was to try to identify responsible genes. Seymour Benzer developed a strategy for identifying genes responsible for traits by exposing fruit flies to mutagenic agents and linking resulting aberrant traits to the mutated gene. In 1971, as a graduate student with Benzer, Konopka pursued this approach to circadian rhythms in fruit flies, creating mutants that were either arrhythmic or exhibited shortened (20 hour) or lengthened (28 hour) rhythms.[9] He traced all these effects to a mutation at a common location on the X chromosome and named the responsible gene *period* (*per*). A few other loci at which mutations altered clock behavior were soon after identified in fruit flies and in fungi[10] and a decade later in hamsters.[11] Initially much of the research focused on carefully describing the behavior of the mutants, including their responses to light pulses. Although there were several attempts to infer the mechanism from the behaviors of the mutants and other clues,[12] these efforts were unsuccessful in providing empirically grounded hypotheses until cloning technology made it possible to study the transcripts of genes and identify their protein products. Using this approach, in 1990 Hardin, Hall, and Rosbash demonstrated daily oscillations in both *per* RNA and the protein PER, and proposed a transcriptional-translational feedback loop mechanism whereby once PER was synthesized and transported back into the nucleus it would suppress its own transcription until it was degraded, after which more PER

7 Shin-Ichi T. Inouye and Hiroshi Kawamura, "Persistence of Circadian Rhythmicity in a Mammalian Hypothalamic „Island" Containing the Suprachiasmatic Nucleus", in: *Proceedings of the National Academy of Sciences (USA)* 76, 1979, pp. 5962-5966.

8 Martin R. Ralph, Russell G. Foster, Fred C. Davis, and Michael Menaker, "Transplanted Suprachiasmatic Nucleus Determines Circadian Period", in: *Science* 247, 4945, 1990, pp. 975-978.

9 Ronald J. Konopka and Seymour Benzer, "Clock Mutants of *Drosophila Melanogaster*", in: *Proceedings of the National Academy of Sciences (USA)* 89, 1971, pp. 2112-2116.

10 Jerry A. Feldman and Marian N. Hoyle, "Isolation of Circandian Clock Mutants of *Neurospora Crasa*", in: *Genetics* 75, 1973, pp. 605-613.

11 Martin R. Ralph and Michael Menaker, "A Mutation of the Circadian System in Golden Hamsters", in: *Science* 241, 1988, pp. 1225-1227.

12 Leland N. Edmunds, *Cellular and Molecular Bases of Biological Clocks: Models and Mechanisms for Circadian Timekeeping.* New York: Springer-Verlag 1988.

could be synthesized (Figure 1).[13] With appropriate delays for the various stages, they hypothesized that this process could generate 24-hour oscillations.

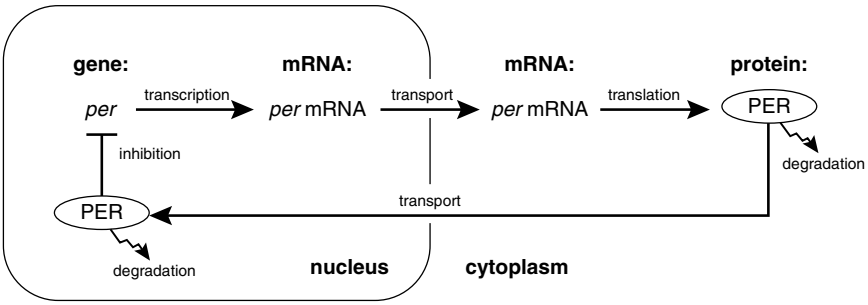


Figure 1. The translation-transcription feedback mechanism proposed by Hardin et al.

In the early 20th century engineers discovered, often to their chagrin, that negative feedback can generate oscillations and mathematically inclined biologists, noting the frequency of oscillatory behavior in living systems, explored the potential of negative feedback to create sustained oscillations. Goodwin, for example, developed a model based on the negative feedback mechanism that Jacob and Monod[14] had proposed for gene regulation in bacteria.[15] In simulations run on an analog computer, he found that he could only generate sustained oscillations when he included at least one non-linear function (involving the Hill coefficient, widely employed in kinetic analyses of biochemical reactions to characterize the number of molecules that must cooperate to achieve inhibition) and even then only when parameters were in restricted ranges.[16] To determine whether the transcription-

13 Paul E. Hardin, Jeffrey C. Hall, and Michael Rosbash, "Feedback of the *Drosophila Period* Gene Product on Circadian Cycling of Its Messenger Rna Levels", in: *Nature* 343, 6258, 1990, pp. 536-540.

14 François Jacob and Jacques Monod, "Genetic Regulatory Systems in the Synthesis of Proteins", in: *Journal of Molecular Biology* 3, 1961, pp. 318-356.

15 Brian C. Goodwin, *Temporal Organization in Cells; A Dynamic Theory of Cellular Control Processes*. London: Academic 1963.

16 In his analog simulations Goodwin reported oscillatory behavior with values as low as 2 or 3 for the Hill coefficient, but shortly afterward Griffith found in digital simulations that undamped oscillations would only occur with values greater than 9, generally recognized as biologically unrealistic: see J. S. Griffith, "Mathematics of Cellular Control Processes I. Negative Feedback to One Gene", in: *Journal of Theoretical Biology* 20, 2, 1968, pp. 202-208. Accordingly, he concluded that negative feedback with a single gene product operating on a gene could never "give rise in practice to undamped oscillations in the concentrations of cellular constituents." Subsequently models, such as those of Goldbeter (discussed below) employ additional nonlinearities elsewhere in the model (e.g., involving the degradation of various components) and so are able to use values of the Hill coefficient that are more biologically realistic.

translation feedback loop proposed by Hardin et al. would be able to generate the phenomenon, Goldbeter elaborated on Goodwin's model. With parameters that he claimed were biologically plausible, Goldbeter's model generated sustained oscillatory behavior.[17]

The research described so far illustrated the combination of tools for decomposition and recomposition in generating an account of a mechanism for circadian rhythms. The mutant research together with cloning techniques allowed researchers to decompose the mechanism, identify an important part – the gene *per* – and characterize an operation in which it engaged – being transcribed into RNA and a protein, both of which oscillated on a 24-hour cycle. This enabled them to recompose the mechanism by proposing a feedback process that could be represented in a diagram. Hardin et al. could verbally describe the behavior such a mechanism might exhibit, but Goldbeter's computational model showed that if the parts operated as Hardin et al. proposed, the mechanism would generate sustained oscillations.

At the same time as Goldbeter was developing his model, other researchers were identifying a host of additional genes in which mutations resulted in altered circadian rhythms and were able to specify the operations in which these figured. For example, by pursuing a strategy similar to Konopka's, Sehgal, Price, Man, and Young identified a second fruit fly gene, which they called *timeless* (*tim*), in which mutations resulted in altered rhythms.[18] In further research they revealed that TIM forms a dimer with PER before entering the nucleus and it is the dimer that figures in inhibiting transcription of both *per* and *tim*.[19] Adopting the same strategy with mice, Vitaterna, King, Chang, Kornhauser, Lowrey, McDonald, Dove, Pinto, Turek, and Takahashi identified a gene they named *Clock* in which mutations resulted in loss of circadian rhythms.[20] Homologues of *Clock* were found in fruit flies, and CLOCK was shown to bind to the promoter of *per* and *tim*. Two homologs of PER, PER1 and PER2, were soon after identified in mice, where they were shown to form dimers not with TIM but with two cryptochromes, CRY1 and CRY2. In short order investigators determined that in mice CLOCK forms a dimer with BMAL1.

17 Albert Goldbeter, "A Model for Circadian Oscillations in the *Drosophila* Period Protein (Per)", in: *Proceedings of the Royal Society of London. B: Biological Sciences* 261, 1362, 1995, pp. 319-324.

18 Amita Sehgal, Jeffrey L. Price, Bernice Man, and Michael W. Young, "Loss of Circadian Behavioral Rhythms and *Per* Rna Oscillations in the *Drosophila* Mutant *Timeless*", in: *Science* 263, 1994, pp. 1603-1606.

19 Leslie B. Vosshall, Jeffrey L. Price, Amita Sehgal, Lino Saez, and Michael W. Young, "Block in Nuclear Localization of *Period* Protein by a Second Clock Mutation, *Timeless*", in: *Science* 263, 5153, 1994, pp. 1606-1609.

20 Martha Hotz Vitaterna, David P. King, Anne-Marie Chang, Jon M. Kornhauser, Phillip L. Lowrey, J. David McDonald, William F. Dove, Lawrence H. Pinto, Fred W. Turek, and Joseph S. Takahashi, "Mutagenesis and Mapping of a Mouse Gene, *Clock*, Essential for Circadian Behavior", in: *Science* 264, 5159, 1994, pp. 719-725.

Another protein, REV-ERBα, was discovered to bind to the promoter of BMAL1 and inhibit its transcription and translation and various kinases were identified as figuring in the phosphorylation of PER and CRY, a factor crucial both in their transport into the nucleus and in their degradation.

The discovery of these additional parts and operations led to new challenges in recomposing the clock. Since each component could be related in one way or another to PER, it was possible to connect them into a common diagram in which the transcription-translation feedback loop involving PER was the central feature. Researchers recognized that there is a second feedback loop in which the action of BMAL1 in activating the production of REV-ERBα is subsequently inhibited when REV-ERBα inhibits the production of BMAL1. Numerous diagrams similar to Figure 2 appeared to illustrate how the various components were thought to be related so as to generate oscillations. However, although one might mentally rehearse the operations portrayed in Figure 1 to show that it might oscillate, this proved harder to do as additional components and feedback loops were introduced. This made it even more important to represent the hypothesized mechanism in computational models to determine how it will behave. In collaboration with Leloup, Goldbeter added terms and equations to his 1995 model to represent both the fruit fly[21] and the mammalian[22] circadian mechanism. In addition to capturing the basic oscillation, Leloup and Goldbeter demonstrated that the components hypothesized to entrain the clock to light-dark cycles could indeed modify the phase of the oscillator in an appropriate manner and that manipulations in the model that correspond to altering components of the clock could generate the patterns of known circadian pathologies such as delayed and advanced sleep phase syndromes.

21 Jean-Christophe Leloup and Albert Goldbeter, "A Model for Circadian Rhythms in *Drosophila* Incorporating the Formation of a Complex between the Per and Tim Proteins", in: *Journal of Biological Rhythms* 13, 1, 1998, pp. 70-87.

22 Jean-Christophe Leloup and Albert Goldbeter, "Modeling the Mammalian Circadian Clock: Sensitivity Analysis and Multiplicity of Oscillatory Mechanisms", in: *Journal of Theoretical Biology* 230, 4, 2004, pp. 541-562.

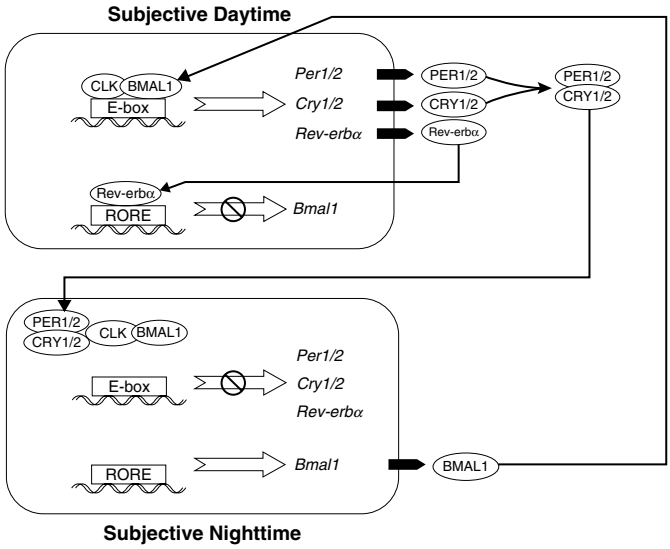


Figure 2. A representation of the mammalian circadian mechanism incorporating many of the additional components that were identified in the 1990s.

3. Systems biological approaches to the oscillator mechanism

The basic research on the circadian oscillator described in the previous section all fit within the framework of molecular biology although the modeling endeavors already foreshadowed the application of the approach of systems biology. Over the last decade the term *systems biology* has been adopted in many domains of biology to signify an approach that focuses on the integration and interaction of large numbers of components giving raise to behaviors that are not readily traced to individual components.[23] Two aspects of systems biology have been particularly important for circadian rhythm research. The first is the introduction of new techniques for identifying large numbers of components that figure in a mechanism (in contrast to the identification of individual parts one at a time as in the genetic research discussed above). For example, a genome wide screen using complementary DNA (cDNA) overexpression assays identified RORα as an activator of BMAL1 transcription that competes with inhibitor REV-ERBα and yields a positive feedback loop.[24] Similar screening techniques revealed numerous addi-

23 Denis Noble, *The Music of Life: Biology Beyond the Genome.* Oxford: Oxford University Press 2006; Hiroaki Kitano (Ed.), *Foundations of Systems Biology.* Cambridge, (Mass.): The MIT Press 2001; Sangdun Choi (Ed.), *Introduction to Systems Biology.* Totowa, NJ: Humana Press 2007.

24 Trey K. Sato, Satchidananda Panda, Loren J. Miraglia, Teresa M. Reyes, Radu D. Rudic, Peter McNamara, Kinnery A. Naik, Garret A. FitzGerald, Steve A. Kay, and John

tional clock components, including various kinases that figure in post-translational modification of proteins. A small interfering RNA screen (siRNA) identified more than 200 genes, many of which figure in different cell-signaling pathways that affect amplitude and period of circadian oscillations.[25] One consequence of this use of systems approaches has been to reveal ways in which the clock mechanism is linked to and affected by other cell functions.

The second contribution is to bring the tools of dynamical systems analyses of mathematical models to bear in understanding mechanisms in which multiple interacting non-linear processes defeat the prospect of understanding the mechanism by tracing out its operations sequentially. Already in his 1995 model Goldbeter pioneered this approach: to show that the model produced sustained oscillations he showed that it generated limit cycle behavior. As I noted, Goldbeter continued this endeavor as new clock components were identified, developing models incorporating all the known constituents of the clock mechanism. While his models generated many features of circadian clock behavior, their very complexity made it difficult to determine which operations in the mechanism were primarily responsible for specific behaviors. Many modelers accordingly prefer to construct reduced models that focus on select components and to manipulate (experiment on) these models to understand what individual components contribute. Accordingly, Smolen, Baxter, and Byrne developed a much reduced model for the fruit fly clock that, for example, did not distinguish PER and TIM and did not incorporate the transport of proteins back into the nucleus (instead incorporating a delay between different operations).[26] After establishing that their model generated appropriate oscillations, they explored whether all components of it were required to do so. By fixing the value for CLOCK concentrations they eliminated the second feedback loop involving REV-ERBα and showed that the feedback of PER and TIM on their own transcription was sufficient (as Goldbeter's first model had suggested). Interestingly, recently Relógio, Westermark, Wallach, Schellenberg, Kramer, and Herzel have reached the opposite conclusion.[27] Their model is somewhat more complex, and incorporates the competition between REV-ERBα and RORα, but is still much simpler than Goldbeter's. When they fixed the variable correspond-

B. Hogenesch, "A Functional Genomics Strategy Reveals Rora as a Component of the Mammalian Circadian Clock", in: *Neuron* 43, 4, 2004, pp. 527-537.

25 Eric E. Zhang, Andrew C. Liu, Tsuyoshi Hirota, Loren J. Miraglia, Genevieve Welch, Pagkapol Y. Pongsawakul, Xianzhong Liu, Ann Atwood, Jon W. Huss, Jeff Janes, Andrew I. Su, John B. Hogenesch, and Steve A. Kay, "A Genome-Wide Rnai Screen for Modifiers of the Circadian Clock in Human Cells", in: *Cell* 139, 1, 2009, pp. 199-210.

26 Paul Smolen, Douglas A. Baxter, and John H. Byrne, "Modeling Circadian Oscillations with Interlocking Positive and Negative Feedback Loops", in: *Journal of Neuroscience* 21, 17, 2001, pp. 6644-6656.

27 Angela Relógio, Pal O. Westermark, Thomas Wallach, Katja Schellenberg, Achim Kramer, and Hanspeter Herzel, "Tuning the Mammalian Circadian Clock: Robust Synergy of Two Loops", in: *PLoS Comput Biol* 7, 12, 2011, pp. e1002309.

ing to the concentration of PER:CRY at its mean value, they found that the loop involving BMAL1 was sufficient for oscillations but when they fixed the variables corresponding to CLOCK:BMAL1 and REV-ERBα to their mean values, rendering CLOCK:BMAL1 into a constitutive inhibitor and REV-ERBα into a constitutive activator, the oscillations in the variables representing PER, CRY, and the PER:CRY dimer were shortened and soon damped out. They concluded that the cycle involving REV-ERBα and RORα was the core mechanism for generating oscillations, and further, since the *Ror*α RNA was almost constant even in the first simulation, that the inhibitor *Rev-Erb*α was the "driving force" in the oscillator.

One possible response to the divergent results of Smolen et al. and Relógio et al. is to dismiss all such modeling efforts as uninformative (since each explicitly makes simplifying assumptions and so deliberately misrepresents the mechanism). But a different response is to view the models as initial steps towards understanding how the mechanism actually works. A crucial further step is to seek ways to link the models back to the actual mechanism and both examine carefully the assumptions each makes, especially in choosing parameters for the models, and to consider what new experiments might be suggested by the models that can be implemented in actual biological preparations. (Although not directly related to the issue of the two feedback loops, Relógio et al. did make new predictions regarding overexpression of *Ror*α and Rev-Erbα that they then confirmed in slice preparation using a *Bmal1*-luciferase reporter.)

I have highlighted two contributions of systems biology to understanding individual oscillators – identifying additional components and experimenting on models to understand how the operations in the mechanism produced the phenomena. These pursuits support each other. One of the results of identifying additional cell constituents that affect clock operation is to show how clock operation is integrated with many other cell activities, including basic metabolism and cell division. Such discoveries make reliance on modeling ever more crucial to understanding how the mechanism will behave in the interactive context of a cell.

4. Systems perspectives at higher levels of organization

At the outset I described how the circadian clock in mammals was initially localized in the SCN. Research on the SCN revealed subpopulations of cells that exhibit different behavior. A basic division was observed between a core region, whose cells express vasoactive intestinal polypeptide (VIP), and a shell region, whose cells express vasopressin.[28] Nonetheless, initially it was plausible to assume that the intracellular oscillator functioned similarly in different cells. However, when Welsh cultured SCN neurons on a multi-electrode array that nonetheless retained

28 Anthony N. van den Pol, "The Hypothalamic Suprachiasmatic Nucleus of Rat: Intrinsic Anatomy", in: *The Journal of Comparative Neurology* 191, 4, 1980, pp. 661-702.

"abundant functional synapses" and recorded from individual neurons, he found that the neurons exhibited a wide variety of phases and periods. Some neurons generated maximal output while others were largely quiescent and their periods ranged from 21.25 to 26.25 hours with a SD of 1.25 hours.[29] Since the SCN as a whole produces a regular output and the variation is eliminated even in explants as long as nearly all the connections are maintained, researchers recognized that communication between neurons is responsible for regularizing the behavior of the individual neurons.[30]

Only computational modeling can illuminate how linking individually variable oscillators into a network could result in each behaving regularly. In a first effort, Gonze, Bernard, Waltermann, Kramer, and Herzel employed Goodwin's model for an oscillator and added terms for the generation of a diffusible compound such as VIP and for the response to its mean concentration and an equation for determining the mean concentration from that generated by each cell.[31] They showed that when the parameter affecting the response to the diffusible compound was set to 0 the model behaved as Welsh's preparation had, but when it was set to 0.5, the oscillators exhibited the synchronization Herzog had found. In their model, Gonze et al. assumed that the network had a fully-connected architecture, one of the modes of organization investigated by graph theorists in the mid-20th century. Two measures are widely employed in analyzing the consequences of network architectures for information flow: characteristic path length and the clustering coefficient. The characteristic path length is the mean of the shortest path between pairs of nodes and reflects how quickly information can be transmitted through the network. The clustering coefficient is the proportion of possible links in local neighborhoods that are actually realized and reflects how much specialized processing can be accomplished by cooperating nodes. Short characteristic path lengths and higher clustering are desirable for information processing and are realized in fully connected networks. However, maintaining complete connectivity between all neurons in a network is metabolically very expensive and so not found in biological systems.

Graph theorists in the mid-20th century also explored two architectures with reduced connections: randomly connected networks and regular lattices. Each only provides one of the valuable characteristics: randomly connected networks exhibit short characteristic path length but low clustering, whereas regular lattices

29 David K. Welsh, Diomedes E. Logothetis, Markus Meister, and Steven M. Reppert, "Individual Neurons Dissociated from Rat Suprachiasmatic Nucleus Express Independently Phased Circadian Firing Rhythms", in: *Neuron* 14, 4, 1995, pp. 697-706.

30 Erik D. Herzog, Sara J. Aton, Rika Numano, Yoshiyuki Sakaki, and Hajime Tei, "Temporal Precision in the Mammalian Circadian System: A Reliable Clock from Less Reliable Neurons", in: *Journal of Biological Rhythms* 19, 1, 2004, pp. 35-46.

31 Didier Gonze, Samuel Bernard, Christian Waltermann, Achim Kramer, and Hanspeter Herzel, "Spontaneous Synchronization of Coupled Circadian Oscillators", in: *Biophysical Journal* 89, 1, 2005, pp. 120-129.

yield high clustering but long characteristic path lengths. However, in 1998 Watts and Strogratz directed attention to a different network architecture. In what they termed "small worlds" most connections are between nearby units, as in regular lattices, but there are a few long-distance connections.[32] The clustering coefficient of such networks closely approximates that of regular lattices, but the characteristic path length is approximately that of a fully connected network. Watts and Strogratz also showed that many real world networks, including biological networks such as the neural network of the nematode worm *Caenorhabditis elegans*, exhibit small-world properties and argued that they could synchronize oscillators nearly as quickly as totally connected networks. Not enough is known of the structure of the SCN to ascertain whether it structurally exhibits the properties of a small world. Instead Vasalou, Herzog, and Henson pursued the strategy of modeling the SCN as a small world and comparing the behavior of the model with the behavior of the SCN.[33] They modeled each neuron using the Leloup and Goldbeter model of the mammalian oscillator modified to include VIP synthesis and set parameter values so that only some of the neurons sustained oscillations when VIP synthesis was suppressed. They organized these into a small world network structure and showed that it would generate synchronization as effectively as a totally connected network. They were also able to capture three other phenomena observed in experimental studies: with VIP (1) the percentage of oscillating neurons in the SCN rises from about 30% to nearly all, (2) the period is extended from approximately 22 to approximately 24 hours, and (3) the variability in periods is largely eliminated.

In these models researchers assumed each cell maintained a given oscillatory pattern except as synchronized with others, but Meeker, Harang, Webb, Welsh, Doyle, Bonnet, Herzog, and Petzold recently employed wavelet analysis which reveals that individual neurons vary in their periodicity, sometimes showing periods greater than 40 hours.[34] To understand what factors accounted for the varying behavior of the individual neurons, Meeker et al. modeled the SCN using a stochastic version of the Leloup and Goldbeter mammalian model and through a series of simulations determined that parameters affecting Bmal1 transcription repression and degradation best accounted for the pattern they observed.

The assumption of near decomposability in traditional mechanistic research makes it difficult for such research to identify, let alone explain, how network or-

32 Duncan Watts and Steven Strogratz, "Collective Dynamics of Small Worlds", in: *Nature* 393, 1998, pp. 440-442.

33 Christina Vasalou, Erik D. Herzog, and Michael A. Henson, "Small-World Network Models of Intercellular Coupling Predict Enhanced Synchronization in the Suprachiasmatic Nucleus", in: *Journal of Biological Rhythms* 24, 3, 2009, pp. 243-254.

34 Kirsten Meeker, Richard Harang, Alexis B. Webb, David K. Welsh, Francis J. Doyle, Guillaume Bonnet, Erik D. Herzog, and Linda R. Petzold, "Wavelet Measurement Suggests Cause of Period Instability in Mammalian Circadian Neurons", in: *Journal of Biological Rhythms* 26, 4, 2011, pp. 353-362.

ganization alters the behavior of individual parts of the mechanism. When complemented by the tools of computational modeling and dynamical systems analyses, though, as posed in accounts of dynamic mechanistic explanation,[35] researchers can both simulate such behavior and begin to understand how the organization of the mechanism explains it.

5. Conclusions

In the 1990s circadian rhythm research made enormous progress in identifying the components of the circadian clock and the operations they performed employing the techniques of genetics and molecular biology. Researchers could recompose the clock in a diagram that showed how the components were related, but to show that by performing the operations attributed to them the mechanism would generate sustained 24-hour oscillations required supplementing these traditional mechanistic approaches with computational modeling approaches developed in systems biology. The need for modeling has grown in the past decade as other approaches from systems biology have revealed more components of cells that affect clock function. As I have illustrated, to begin to understand what parts of the mechanism are responsible for sustained oscillations, researchers resorted to developing simplified models and performing manipulations on them. In addition to facing these challenges in understanding the intracellular mechanism, researchers also came to recognize that the oscillators are incorporated in networks and that only as part of the network do they generate sustained circadian oscillations. Again, to understand how coupling into networks alters the behaviors of the components and generates regular behavior requires modeling and systems analysis. This need to turn to systems biological approaches is itself driven by discoveries about the mechanism responsible for circadian rhythms.

Department of Philosophy and Center for Chronobiology
University of California, San Diego
La Jolla, CA 92093-0119
USA
bill@mechanism.ucsd.edu

35 See Bechtel, *op. cit.* and Bechtel and Abrahamsen, *op. cit.*

Olaf Wolkenhauer and Jan-Hendrik Hofmeyr

Interdisciplinarity as both Necessity and Hurdle for Progress in the Life Sciences

Abstract

The ability to sequence the genome of entire organisms has produced a fundamental change in the scientific practice of the life sciences. With the Omics revolution, biologists working with cellular systems have become dependent on the support of and collaboration with other disciplines. Following the identification and characterization of cellular components in the context of bioinformatics, the focus has shifted in recent years to the study of mechanisms that determine the functioning of cells in terms of gene regulatory networks, signal transduction and metabolic pathways. This shift of focus towards an understanding of functional activity and therefore towards cellular processes required methodologies from systems theory and thus expertise from other fields than computer science and physics. Since then, the term 'systems biology' has become associated with an interdisciplinary approach that realizes a practice of data-driven modelling and model-driven experimentation. With systems biology, mathematical models have become a central element in the formulation of biological arguments and as a consequence, a new quality of interdisciplinary collaboration has become necessary. The "modeller" or "theoretician" no longer plays a simple supportive role. Instead, the construction and analyses of the models require both – the "experimentalist" and "modeller" to meet at "eye level", pursue a common question, and rely upon each other. The present text discusses the practice of systems biology with respect to the hurdles and opportunities provided by interdisciplinary collaborations in this field. The main conclusion is that truly interdisciplinary collaborative efforts are a necessity for progress in the life sciences but these efforts are hampered by academic structures and practices that prevent these projects from succeeding.

1. The emergence of systems biology

The ability to sequence the genomes of organisms has produced a fundamental change in the scientific practice of the life sciences. Genome projects have generated large-scale data sets, which required databases to store information about sequences, structures, and auxiliary information about the gene and proteins in question. In addition to the computational infrastructure that stores the data and the provision of interfaces to access the information, tools and algorithms were

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_18,

needed to analyse the data. To this end, predominantly statistical and machine learning techniques were employed, thereby attracting computer scientists and physicists to the life sciences. While computer scientists have often cast themselves in a supportive, software-developing role in which biological questions are of secondary interest, the skilled application of tools and algorithms to answer biological question has turned many biologists into "bioinformaticians". One possible explanation for the success of physicists in the biological sciences may be seen in their training – they are competent in mathematical modelling, not afraid of theory but at the same time they do not mind "getting their hands dirty" with experimental data, which they know to process with statistical tools. In systems biology, a similar argument can be made about control engineers, who are trained to combine mathematical modelling with experimental data and "real-world" problems.

With the genomics revolution, biologists have thus become dependent on the support of and collaboration with other disciplines. Genomics and bioinformatics has focussed on the identification and molecular characterization of cellular components. It quickly became apparent that from the components themselves one cannot fully understand their function. This triggered a shift of focus towards the study of interactions and an understanding of mechanisms that underly cell function (e.g. cell growth, proliferation, differentiation and apoptosis). The study of functional activity as processes that are realized by gene regulatory networks, signal transduction pathways and metabolic networks requires techniques to model dynamical systems.

The fact that cell functions are driven by spatio-temporal processes is crucial for the emergence of systems biology. While statistical and computational techniques (including machine learning) took centre stage in bioinformatics, the shift of focus towards an understanding of functional activity and processes required methodologies from systems theory. This need introduced many (control) engineers to the biological sciences. The use of systems-theoretic approaches in molecular and cell biology, mostly focussing on intracellular pathways and networks, has now become an active area of research under the umbrella of 'systems biology'. The practice of systems biology is characterized by a close integration of "theory" and "experiment", of data-driven modelling and model-driven experimentation. The role of mathematical models is changing from a supportive to a central role in formulating and arguing a biological hypothesis. As a consequence, a new quality of interdisciplinary collaboration has become necessary. The "modeller" or "theoretician" no longer plays a simple supportive role in which a deeper understanding of the biological context is secondary. Instead, the "experimentalist" and "modeller" have to meet at "eye level", pursue a common question, and rely upon each other for their mutual success. The remainder of the present text discusses the practice of systems biology with respect to the hurdles and opportunities for interdisciplinary approaches.

Truly interdisciplinary, large-scale, and multinational projects are essential for progress in the life sciences. The complexity of cells and systems made up of cells

does not leave us a choice. We shall here argue that the biggest hurdle for progress may not be funding or technical limitations, but the personal relationships and dependence of members in interdisciplinary teams. In many instances the risk of failure in a project is *not* related to the science or scientific approach, but is more often a consequence of personal problems between project leaders, often as a consequence of the current academic system and how this system hinders interdisciplinarity.

2. True interdisciplinarity

Any definition of interdisciplinarity already harbours some difficulties. Experts differentiate between trans-, inter- and cross-disciplinarity. What we will refer to in the present text, focussing on the practice of systems biology, is interdisciplinarity understood as a means to answer questions by teams of experts from different disciplines *and* which could otherwise not be answered. We therefore speak about the collaboration of at least two experts from different fields of research. In a truly interdisciplinary project, the team members meet "at eye-level", sharing a passion for *one and the same* research question – with either of them having only a small or no chance of succeeding on their own. This is the crucial point – a blessing and a curse at the same time. Because all team members share an interest in the same question but approach it from different angles, this collaboration has the greatest potential for finding an answer or novel solution. At the same time, however, the participants in such a truly interdisciplinary team will depend on each other, on their ability and compatibility.

Interdisciplinarity is a key element in large-scale research projects. However, the inevitable loss of independence in a truly interdisciplinary project provides an enormous challenge to the realization of large-scale research projects in the life sciences. At present, many projects in the life sciences that may be considered "large scale" efforts are more often than not characterized by redundancy and a strong degree of independence of the partners. In such projects each partner contributes a piece to a puzzle but the search for and description of that piece is something the partner can do fairly independent of other partners and with only infrequent interactions. In such large-scale projects, data and models are not integrated at the level of experiments; instead, the results of subprojects, the interpretation of individual works are being integrated at an intellectual level, through discourse and joint publications.

The genome projects are also examples of large-scale collaborative efforts in the life sciences but in these projects the dependency of partners is limited. Here the costs for the hardware and infrastructure, the desire to conduct a comprehensive study, as well as the wish to share the data amongst a large group of users are the main motivations for collaboration.

Health research provides an example of an area where there is an obvious need for integration of research efforts, not just between groups or across disciplinary boundaries, but also across countries. In recent years, funders of research in the life sciences have realized the importance of interdisciplinary research and have established a large number of programmes to promote interdisciplinary collaborations. Systems biology and systems biology approaches have emerged in this period of unprecedented opportunities for interdisciplinary projects. The complexity of cellular systems makes a joint collaborative, truly interdisciplinary effort necessary. As will be discussed below, this however requires that the environment, including the academic structures, are supportive of such efforts. The present situation is one in which large-scale projects do not pursue an integration of results at the level of experiments. This is because the coordination of such projects would require an elaborate strategy and top-down steering to ensure that the extra effort required is not a hurdle.

From a scientific point of view, the development towards truly interdisciplinary projects in biology and medicine may be seen as necessary but it should also be recognized that the scientific effort has to be preceded and anticipated by an enormous effort of the funding bodies and academic system. Interdisciplinary projects not only imply an additional effort by the scientists but also a greater effort in their administration, evaluation and coordination. For the purpose of the present essay, we shall however focus on our experiences as scientists and what we consider the biggest hurdle for progress in systems biology, respectively systems medicine.

3. Apparent interdisciplinarity

New funding programmes that support interdisciplinary efforts in systems biology are a temptation to anyone who seeks funding for his/her research. The many existing interpretations of the nature of systems biology are, in part, also a reflection of the creativity of scientists to attract funding. The re-labelling of one's own work as "systems biology" without ever changing the scientific approach presents a serious threat to progress in the life sciences. The need for interdisciplinary approaches, like systems approaches and mathematical and computational modelling, is a consequence of and response to biological complexity. The misuse of the term "systems biology" for pseudo-collaborative projects undermines the real added value of interdisciplinarity. A point in case is the boundary between bioinformatics and systems biology. Bioinformatics approaches are characterized by the use of algorithms, say for sequence analysis, structure prediction and the use of databases, machine learning techniques and statistical techniques. There is a focus on macromolecules and if networks are considered, then temporal aspects do not play a role. In contrast, systems biology has emerged from the realization that cell

functions are driven by networks of genes and proteins interacting in time and space, leading to the view that the functioning of cells is an intrinsically dynamic phenomenon. Once one accepts that a cell function, say apoptosis, is a nonlinear dynamical process, then the theory of dynamical systems should or must enter the scene. A subtle but important difference between bioinformatics and systems biology is thus the perspective, that is, the focus on individual components vs. dynamical networks. For this reason different people are attracted to the fields: in bioinformatics, mostly computer scientists and biologists can be found, while in systems biology the theory of dynamical systems is more important and hence engineers and applied mathematicians will feel more at home. However, with the broad range of problems and due to the fact that many approaches are complementary, it is difficult to draw clear boundaries.

The point is that biological complexity forces us to expand our set of tools, and sometimes a change in how we pursue a problem become necessary. To ensure that such change is implemented, may require some top-down steering to put pressure upon scientists to change their practice. While one would naively imagine scientists to choose whatever is best for answering their scientific problem, in reality decisions are more closely linked to administrative and formal requirements for career progression. There is also an inherent resistance to change if it is accompanied by an additional effort. As will be discussed below, in systems biology the collaboration of experimentalists with modellers usually implies more costly and more time-consuming experiments. Mechanistic models of cell functions require sufficiently rich, quantitative datasets, the need for replicates and increased precision to quantify the uncertainty in experiments; this can strain the relationship in any interdisciplinary partnership. Mathematical modelling represents however not only a natural language with which to integrate data at various levels, the theory of dynamical systems becomes a necessity when dealing with complex dynamical phenomena. Conventional models of medical and biological explanation rely primarily on verbal reasoning and are only suited for dealing with mechanisms that involve small numbers of components and short chains of causality. The value of modelling is then that is necessitates the statement of explicit hypotheses, a process which often enhances comprehension of the biological system and can uncover critical points where understanding is lacking. Simulations can then reveal hidden patterns and/or counter-intuitive mechanisms. Theoretical thinking and mathematical modelling thus constitute powerful tools to integrate and make sense of biological and clinical information being generated and, more importantly, to generate new hypotheses that can then be tested in the laboratory.

Biomedicine is an area in which the need for truly interdisciplinary and integrative large-scale efforts is most obvious. Many diseases are spatio-temporal phenomena that occur across multiple levels of structural and functional organization of the human body. Understanding diseases requires an integration of data from the cellular level to the physiology of an organ. Another dimension is the need for an integration of experimental systems, merging results from studies on

cell cultures, genetic/mouse models and patient data. The data themselves can be generated with a range of technologies, each of which is often a specialization with its own community, journals etc. At present there is no experience with such comprehensive, large-scale projects of an interdisciplinary nature.

Only "truly interdisciplinary" projects are likely to provide a high degree of innovation, a "more valuable" publication output (publications that would otherwise not have been achieved and which are published in journals with a higher impact). Most importantly, truly interdisciplinary projects increase the chance of solving complex problems. Research funders and decision-makers should care (more) about recognizing "real" interdisciplinarity. Freeloaders (as reputed as they might be) should be dropped.

4. Learning from physics

In September 2008, after decades of work and bringing together thousands of scientists from hundreds of institutes, universities and laboratories from more than eighty countries, a particle accelerator called the "Large Hadron Collider" or LHC was launched. The project cost some billions of Euros and is dedicated to the question of whether there is such thing as the Higgs boson – an elementary, hypothetical particle. What this project demonstrates is a culture in which large teams collaborate on a joint project and in which many subprojects are mutually dependent. Physicists dare to make their own success dependent upon the other project partners, technicians and designers of the devices. These collaborative efforts are born out of a necessity dictated by the complexity of the problem at hand. Besides an organizational and communication structure, such projects require persistence and a lot of money, too. Physicists have thus succeeded in convincing the general public and politicians of the importance of their goals.

Comparing large-scale research projects in physics with those in the life sciences, the difference becomes apparent. In the European Union, the largest projects in health research are funded with a maximum of 12 million Euros – this is less than the costs to repair the particle accelerator that broke down right after its launch in September 2008. Projects in the life sciences are usually funded for three years, very rarely for more than five years. Everyone who wants to initiate a comprehensive, multinational disease research project will realize that 12 million Euros is not a lot of money to develop new drugs and therapies – certainly not in three years. Furthermore, existing large-scale projects in the life sciences, on a national and international level, are usually designed in a way that the subprojects are pretty much independent of each other – the risks and the fear of failure is perceived as too high. This is true for biologists and biomedical scientists themselves as well as for the cooperation with theoreticians. No one would consider research on diseases less important and less complex than the search for the Higgs-boson

and yet this is what the current practice implies: we lack a realistic strategy and approach to tackle diseases by developing a culture for truly interdisciplinary large-scale projects in the life sciences.

The technological developments in the life sciences make it necessary for computer scientists, mathematicians, physicists and engineers to get involved. Not only the frequently quoted "flood of data", but the complexity of the processes looked at, create new areas research, among which systems biology and synthetic biology have generated considerable interest. The fact that cell functions are non-linear dynamic processes means that they cannot be analysed by common sense and intuition – as highly developed these faculties might be. It therefore becomes necessary for "modellers" to get involved, to use mathematical modelling as an extension of common sense into the realm of complex systems. This requires a new quality of cooperation, starting from the design of the experiments and ending with the interpretation of data. A partnership like this comes along with some potential for trouble, further discussed below.

5. Systems biology should not be a discipline but an "approach"

No one will doubt that understanding any disease is less complex than the proof for the existence of the Higgs boson that the physicists are striving for. A comprehensive disease project requires the combination of a range of technologies to generate data, the comparison of different experimental systems to compare the results and the integration of results over a wide range of spatio-temporal scales – from the molecular and subcellular level to the physiology of an organ. The large quantities of data and their heterogeneous sources motivate the cooperation with bioinformaticians, while the nonlinearity and the dynamical aspects of cellular phenomena requires systems theoretical approaches. In biomathematics and theoretical biology, groups working on a theoretical basis have been inspired by biology for decades. However, theory and experiment never really intertwined. Now, systems biology gives reason for hope to develop models from experimental data and to use models for the design of new experiments.

Systems biology should not to be understood as a new discipline, but as a new *approach* to examine complex cellular systems. Mathematical methods are used as tools to support common sense and the excellent intuition of an experimentalist in analyzing non-linear, dynamical processes. The process of modelling in itself, and the discussion about what to measure and how, are valuable and help to formulate hypotheses and new experiments with which they can be tested. It is exactly this dialogue that benefits from the different views and different training of the scientists involved. This dialogue is also a reason why this interdisciplinary approach is so exciting and inspiring.

Because systems biology is not a discipline, there are no "systems biologists" either. There are medical scientists, biologists, physicists, mathematicians, control engineers and computer scientists realizing a systems biology approach by their cooperation. To make this work, "only" three things are necessary: [i] specialists from different fields of research, [ii] mutual respect and a basic appreciation of each others work, and [iii] the interest in a joint or shared scientific question that will be solved by meeting one another at "eye-level". This list suggests that the appeal and the risks of interdisciplinary research depend on interpersonal, even psychological factors.

One of us took part in initiating the first international journal on systems biology and fought vehemently for a distinction between bioinformatics and systems biology (albeit always pointing out their complementarity). Unfortunately, it is still necessary to promote systems biology separately to avoid true interdisciplinarity ending prematurely and freeloaders using it as a "buzzword". If systems biology prevails as a new view, a new approach – mathematical modelling being accepted as integral part in answering biological questions – we would not mind if the term "systems biology" disappears as a research field, in the names of departments, or in the names of research institutes. The goal is to answer biological and biomedical questions and recognizing the complexity of cellular systems; this requires a change of practice. The notions of 'systems biology' or 'systems medicine' serve as a vehicle to induce this change.

6. Missing the wood for the trees

To solve important, exciting, scientific questions, generalist-specialists are needed: specialists who dare to look over the edge. This is against the idea that the solution of a complex problem requires specialists only, that is, scientists who spend all of their time concentrating on one single question.

If we are serious about the investigation of diseases, the promotion of true interdisciplinarity will be necessary. The large number of funding programmes available all over the world should provide strong encouragement for interdisciplinary research. However, researchers have to ask themselves at which point they should give up specialization. We would recommend starting to broaden their horizon with the masters degree at the earliest. Interdisciplinarity thrives on the encounter of different views and expertise. Assuming that training in the individual disciplines spans several years, the expertise of two different areas can be hardly reconciled in one person. Specialization is a recipe for success, in nature with plants and animals, as in economies and science. At the beginning of a career it is important to become an expert in a single discipline, to distinguish oneself from the big crowd in order to win the race for university positions and to successfully obtain research funds.

A prime example in which a high degree of specialization and individual effort appears to be common is mathematics. In mathematics, extreme forms of specialization are considered as necessary precondition to succeed. A look at the mathematical highlights of the last few years reveals another facet, though. Although the solution of the Poincaré conjuncture by the Russian mathematician Grigori Yakovlevich Perelman was the result of a focused work during several years by a single person, it was still necessary to combine the results of a range of mathematical research areas to prove it. Geniuses often appear as experts by means of extreme specialization. If one takes a closer look, however, they quite often reveal themselves as generalists in their field of specialization. Fermat's Theorem was supposed to be one of the biggest mathematical problems that could not be solved for several centuries, until Andrew Wiles found the proof in 1995 – after decades of work. His proof, however, compiled and created results in a large number of areas in mathematics. Andrew Wiles is a generalist-specialist, as were Albert Einstein and Leonardo da Vinci. Linus Pauling and Max Delbrück are renowned examples from the life sciences. For Perelman and Wiles, it was necessary to have an extraordinary command of many areas of mathematics. Many curricula vitae of great scientists prove the fact that the look beyond the boundaries of their own specialty does not do any harm.

Truly interdisciplinary large-scale research efforts requires both: the combination of specialists from different disciplines and generalist-specialists. The question for how we can progress in the life sciences has thus also consequences for training researchers.

7. Science evaluation as a threat for interdisciplinarity

With the ever finer branching of the sciences into new disciplines, we must be careful not to miss the wood for the trees. The wood is the nature of complex systems. The complexity of nature makes an interdisciplinary, e.g., systems biological approach, absolutely necessary. For funders such interdisciplinarity must have a high priority and because of the risks involved special attention is required.

But also for universities, "true" interdisciplinary harbours an important potential for success. Interdisciplinarity contributes towards the creation of "critical mass" when it comes to attracting grants for large(r) research projects. Especially for small universities, in which departments tend to be small too, interdisciplinarity is a mechanism to build competitive teams, to have success with larger proposals and to increase international visibility. The trend towards dissolving disciplinary boundaries could be used as an opportunity because in small universities there often are ways to shortcut physical and communication obstructions.

The big funders for research, like the National Institutes of Health (NIH), the European Union and many national funding bodies, have recognized that inter-

disciplinary is necessary but also that the initiation of such collaborations has its problems. In particular the spatial separation, that is, the opportunity for researchers to meet and get to know each other, and the often very different working languages and cultures, can be a hurdle. What funders have recognized is also true for universities: disciplinary boundaries must be overcome systematically. This includes the creation of opportunities during which professors, PhD students and postdocs from different areas can get to know each other. Such get-togethers however have to be actively organized and moderated – the usual form of seminars are not sufficient.

Once new project partners have met and a collaboration has been established, the spatial separation of laboratories can be overcome with, for example, videoconferences that are already common practice in international projects. The common research problem, which is equally exciting to all partners, should make it easy to overcome such practical problems. Another, much bigger problem lurks in the publication of results. Two difficulties come together here: the authorship and the often very different cultures in judging the contributions in the list of authors. If, for example, two professors – one from an experimental group and one theoretical group – collaborate, one can end up with two more postdocs and two more PhD students in the list of authors. How does one rank the names? With the usual project duration of three to four years, it will not be easy to generate enough manuscripts to keep everyone happy, to ensure everyone gets the credit (s)he deserves through an appropriate position in the list of authors. Ideally, the collaboration leads to publications in journals from both fields. Theoretically one could then increase the overall "output". In practice this usually looks different.

In biology and medicine the impact factor of journals plays an important role for in career development. While for most researchers in the medical sciences an impact factor of over 10 is aimed at, for theoretical and mathematical journals the impact factors are far lower, for various reasons to do with the different cultures in these fields. The judgement of interdisciplinary grant applications is often done in relation to the applicant's publications and impact factors of the journals under consideration. At present there is a lack of understanding and appreciation for the different citation cultures and one would expect that various project ideas suffer from poor judgement of the reviewers. Because it should be quality over quantity, it has become common practice to consider the number of citations a paper has received. The Hirsch-(h)-index is very popular and easy to determine for any scientist on the Internet. With all these efforts to evaluate, to quantify, we scientists have accepted the situation in which our efforts and work is reduced to a single number! It is impossible to imagine this for any other part of society, but in science many decisions are taken without a closer look at and discussion of someone's CV. Instead, formulas and counting and indices are used. Encouraging interdisciplinary research requires a strategy and academic structures that avoid pitfalls such as those described here.

8. Summary and conclusions

In summary, a big enemy of interdisciplinarity, and thereby the biggest hindrance to the solution of important scientific questions, is (i) factors in interpersonal relations, (ii) the judgement of authorships in joint publications and (iii) the quantitative evaluation of scientific performance with formulas and indices. Other problems, like physical separation of groups or finding a common language, can, once the collaboration is initiated, be overcome.

There is however no doubt that interdisciplinarity is necessary – not only for science and the solution of important problems, but also for universities to build critical mass. For the young scientist who has specialized during his training, an interdisciplinary orientation offers opportunities to enrich personal experience and to build a career on a broader basis with possibly more opportunities and choices.

With interdisciplinarity comes, above all, the requirement for a greater effort and an increased risk for failure. True interdisciplinarity requires a longer time frame to realize preliminary results; it is hard, or often impossible, to complete a doctoral thesis within three years and the more "leaders" are involved, the greater the potential for conflicts.

There are thus, on the surface, few reasons speaking *for* interdisciplinarity, but those are after all most important. As universities realize the importance of interdisciplinary collaborations to generate the required critical mass for grant applications, interdisciplinarity also provides opportunities for young scientists. Above all, the problems we are trying to solve depend on graduates, postdocs and academics deciding – after years of intense specialization – to take a look beyond their own field.

Compared to the established disciplines, interdisciplinarity is an extreme sport, which requires most of all persistence, risk taking and a long-term effort. Extreme sports, like the practice of systems biology, involves going from one failure to the next without loosing enthusiasm, thereby however pushing the boundaries of what can be achieved. Extreme sports are not for everyone but those who are made for it derive a great deal of satisfaction from it, achieving things that would otherwisenot be possible.

Olaf Wolkenhauer
Department of Systems Biology & Bioinformatics
University of Rostock
18051 Rostock
Germany

Stellenbosch Institute for Advanced Study (STIAS)
Wallenberg Research Centre
at Stellenbosch University
Stellenbosch
South Africa
olaf.wolkenhauer@uni-rostock.de

Jan-Hendrik Hofmeyr
Centre for Studies in Complexity
and Department of Biochemistry
University of Stellenbosch
Stellenbosch
South Africa
jhsh@sun.ac.za

Team C
The Sciences of the Artificial vs. the Cultural and Social Sciences

Amparo Gómez

Archaeology and Scientific Explanation: Naturalism, Interpretivism and "A Third Way"[1]

Abstract

The explanation-understanding controversy has been a main topic of archaeological methodology since the mid 19th century. The arguments for explanation were dominant throughout much of the 20th century within the empiricist and post-empiricist approaches. However, towards the end, understanding approaches were widely adopted by archaeologists, due to the prevalence gained by the interpretive turn in both hermeneutics and post-modern radical version. The aim of this paper is to review the less radical positions within the interpretive turn, that is, the hermeneutical thesis about understanding, and to examine the possibility of convergence between them and post-empiricist approaches on explanation.

1. Contextualizing the explanation-understanding debate in archaeology

Archaeologists have paid a strong attention to the philosophy of science in the search for epistemic and methodological grounds for their discipline. Archaeologists have not been limited to closely follow the debates and arguments of philosophers, but they have developed an interesting epistemic and methodological reflection within archaeology realm. This reflection has been so important that, to a large extent, the evolution of archaeological approaches is the result of the evolution of the philosophical perspectives in the discipline. As Wiley recalls, the training of young archaeologists included philosophy of science, as well as learning specific research techniques.[2]

How archaeologists should explain and, therefore, which is the appropriate model of explanation is one of the central topics of philosophical discussion in archaeology. This discussion has been posed in the context of leading philosophical schools: logical empiricism, structuralism, systems theory, Marxism, post-empiricism, hermeneutics and post-modernism. The above list gives an idea of the dif-

1 This paper has been written thanks to the support of the Spanish Ministry of Science and Innovation research project FFI2009-09483. I am very grateful to Wenceslao J. Gonzalez for his insightful comments and suggestions on earlier drafts of this paper.

2 Alison Wylie, *Thinking from Things, Essays in the Philosophy of Archaeology*. Berkeley: University of California Press 2002, p. XII.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_19, © Springer Science+Business Media Dordrecht 2013

ficulty entailed in addressing a debate that combines the usual philosophical topics with the specific issues of archaeological explanation. The difficulty increases if it is also taken into consideration that archaeology has been understood as history of the facts of the past, natural science and cultural history.

The discussion about the appropriate model of archaeological explanation has had one of the major subjects in the explanation-understanding debate. What is at stake in this debate is the nature of archaeological explanation, and thus of archaeological knowledge. Significant methodological and epistemic views underlying this debate make it fundamental for assigning archaeology to one or the other of the perspectives indicated above. On the other hand, the arguments presented and their impact have considerable interest to the general philosophical discussion on explanation and understanding, and also for the inquiry into the boundaries between social sciences and cultural sciences. However, surprisingly, archaeology has not been considered in philosophical discussions and it has been situated in a separate field, cultivated basically by archaeologists themselves; something that should be corrected, given the fertility of the philosophical discussion that takes place within it.

The explanation-understanding debate in archeology has its origins in the scientificist turn that this discipline took in the 50s in response to a first stage of the archaeological knowledge based on empirical data and their interpretation. Archaeology was established in the mid 19th century and was understood as a history with imprecise narratives of the influences between cultures.

This traditional approach was questioned by North American and British archaeologists, who were persuaded that archaeology should be a science like the natural sciences, able to establish genuine scientific explanations of objective facts. This new view was developed mainly in the 60s with the work of a group of young archaeologists, headed by Binford.[3] This approach was called *New Archaeology* and also *Processual Archaeology*.

The New Archaeologists embraced neo-positivism as the genuine scientific philosophy. The scientific explanation was the Deductive-Nomological (D-N) model, and archaeological explanations should be based on confirmed universal laws. Hence it was essential to confirm the universal laws of cultural and historic processes of the past through the data found in the present excavations. In this context it was rejected any explanation different from the D-N model, and particularly any use of interpretation and understanding of archaeological data.

Despite this radical approach, processual archaeologist soon saw that was impossible to avoid the recourse to interpretation in both the explanation and the confirmation of the laws. One of the main difficulties was that some degree of interpretation of the data was necessary in order to test the universal hypotheses; in the case of explanation, that the connection between hypothesis and data was not

3 See Lewis Binford, "Archaeology as Anthropology", in: *American Antiquity* 28, 1962, pp. 217-225. Lewis Binford and Sally R. Binford (Eds.), *New Perspectives in Archaeology.* Chicago: Aldine 1968.

deductive because it implied interpretation of the recorded data in accordance with said hypothesis. Binford, one of the most significant processual archaeologists, accepted the existence of auxiliary hypothesis connecting data with the hypothesis to be confirmed through interpretations of the meaning of these data,[4] because "the facts of the records do not have a clear and unambiguous meaning".[5] In the late 70s and in the 80s New Archaeologists had to face the proposal made by Kuhn and other authors of Post-empiricist Philosophy. Processualists admitted the thesis of the theory laden of observation, and the same Binford stated:

> objectivity was not attainable either inductively or deductively (…) Archaeological knowledge of the past is totally dependent upon the *meanings* that archaeologists give to observations on the archaeological record (…) there is not independent grounds for proving a hypothesis.[6]

The methodological approaches evolved and diversified, and archaeologists followed two main pathways in explanation: a) to defend the scientific explanation, but not the DN model, and b) explanation was rejected in favor of understanding, insofar as many archaeologists were increasingly pessimistic about processual archaeology, even in its pot-empiricist version. They denied that archaeology should be a science like the natural sciences, and turned their attention back to considering it as the history of the cultures of the past, and as a humanist discipline. This new approach was called *post-processual archaeology* (given its rejection of processual archaeology).

This paper focuses on interpretative turn in post-processual archaeology. It will be centered in the efforts made by archaeologists to avoid understanding relativism and subjectivism by articulating some methodological requirements from the post-empiricist archaeology in the interpretative field, – rather than from traditional hermeneutical resources. The aim of the analysis is to explore possible convergences between explanation and understanding in this context, as a first step towards a third way for a classic debate in archeology (and social sciences) with the purpose of exploring new perspectives on its settlement.

2. Post-processual archaeology and the interpretative turn

Post-processual archaeology emerged in the 80s and it was consolidated in the 90s. This approach took its inspiration from a range of schools of thought, hermeneutics, symbolic and structuralist trends, neo-Marxism, post-structuralist and

4 See Binford, *An Archaeological Perspective*. New York: Seminar Press 1972.

5 Binford, *Nunamiut Ethnoarchaeology*. New York: Academic Press 1978, p. I.

6 Lewis Binford and Jeremy Sabloff, "Paradigms, Systematics and Archaeology", in: *Journal of Anthropological Research* 38, 1982, pp. 138-139. (*mine*).

post-modern thought.[7] Many different types of research questions have their place in post-processual archaeology: gender, power, symbolism, ritual action, personal identity, nationalism, and so on.

One of the most prominent approaches in post-processual archeology has been the *interpretative turn.* The focus of interpretative archaeology is that the past is made intelligible through interpretation and understanding of the actions of situated individuals who produced the material culture, whose remains are studied by archaeologist through the present data records. Thus, interpretative archaeologists do not accept that societies are made up of a series of underlying mechanisms, process, patterns or forces that determine human behaviour, and neither do they consider collective or group behaviour to be the essential. Therefore, archaeologists should not inquire into these fields to understand the cultures of the past, but in the intentions, thoughts and reasons of people to act as they did it.

General theories are rejected to the extent that each culture is a specific case and it must be studied as such (they are opposed to the great theories like the eco-materialists theories). The variability of material culture cannot be explained in terms of laws, generalisations, models or functions, but interpreting and understanding human action and activity, just what the processualists considered epiphenomena without any explanatory capacity.

Understanding is considered the art of understanding meaning, of making it comprehensible, and data record meaning is achieved through interpretation. The most basic level of access to the world is the interpretive one, since what it is observed means something insofar as it is interpreted it, and then can be understood. The main argument of interpretative archaeology is, that there is no definitive knowledge of the past, and "no single methodology can reveal it to us".[8]

This approach has to deal with some questions that are transferred from hermeneutics to archaeology: the problem from where it is made the interpretation,

7 Many are the authors in each tendency; some of the most prominent are: in hermeneutics particularly Ian Hodder, *Reading the Past: Currents Approaches to Interpretation in Archaeology*. Cambridge: Cambridge University Press 1986. Hodder, "Interpretative Archaeology and Its Role", in: *American Antiquity* 56, 1991, pp. 7-18; in structuralism, André Leroi-Gourhan, *L'art parietal: langage de la préhistoire.* Paris: Éditions Jérôme Millon 2009. Robert W. Preucel, "The Postprocessual Condition", in: *Journal of Archaeological Research* 3, 1995, pp. 147-175; in Neo-Marxism, Mark P. Leone (Ed.), *Historical Archaeologies of Capitalism*. New York: Kluwer 1999. Brucer Trigger, "Hyperrelativism, Responsibility, and the Social Sciences", in: *Canadian Review of Sociology and Anthropology* 26, 1989, pp. 776-797; and in post-modern archaeology, Christopher Tilley, *Material Culture and the Texts: The Art of Ambiguity.* London: Routledge 1991 or John Bintliff, "Postmodernism, Rhetoric and Scholasticism at TAG: The Current State of British Archaeology", in: *Antiquity* 65, 1995, pp. 274-278.

8 Julian Thomas, "Introduction: The Polarities of Post-processual Archaeology", in: Julian Thomas (Ed.), *Interpretative Archaeology. A Reader*. London: Continuum International Publishing Group 2002, p. 3.

(just from the present, or the past has any weight?) and so, relativism and subjectivism problems.

Hodder, the most influential author of interpretative turn in archaeology addresses these problems following mainly to Collingwood, but also to Dilthey. Hodder believes that archaeology maintains close links with history and historicism. Material culture is significantly constituted and it is the result of deliberate actions by individuals whose intentions must interpreted in order to understand this culture. If archaeologists have to understand the past, they must pay attention to what Collingwood described as "insides of actions" (despite the inferential distance that they may have), that is, the thoughts and intentions behind the events of the past. Of course the outside of the events are the first discovery, but as the event really important is an action, it is necessary "to get at the subjective meanings, at the inside of events".[9] Hodder subjectivism is an attempt to "interpret the evidence primarily in terms of its internal relations rather than in terms of outside knowledge".[10] To this end, archaeologists must explore research strategies that make it possible to understand cultures as significant products that encode subjective meanings. This kind of understanding involves both the past and the present from where interpretations are made; notions of the past and present can enter into a dialogue, but the past is interpreted in terms of the present.

The answer given to the problem of subjectivism by Hodder, during the 80s, is the same that gave Collingwood and Dilthey with the introduction of "an objective mind capable of bridging the distance between the intentional meanings of past individuals, permanently fixed life expressions and our own understanding in the present".[11] But in two works published in the 90s Hodder takes distance of Collingwood and deals with Gadamer, Ricoeur and Habermas's hermeneutics. In these works Hodder does some comments on method and objectivity in order to introduce a critical and political dimension in hermeneutic archaeology, regarding the incorporation of "other voices" in the interpretation of the past. From this point of view, he argues the critical role of the data, claiming that, "the moment of critique in the hermeneutic processes is the interaction with data to produce 'possible worlds'".[12] Therefore he defends certain room for objectivity with expressions as "a guarded objectivity of the past",[13] "the organized material has an independence",[14] or "material culture as excavate by archeologists is dif-

9 Hodder, *Reading the Past: Currents Approaches to Interpretation in Archaeology*, *op. cit*., p. 79.

10 Hodder, *The Domestication of Europe.* Oxford: Blackwell 1990, p. 21.

11 Harald Johnsen and Bjørnar Olsen, "Hermeneutics and Archaeology. On the Philosophy of Contextual Archaeology", in: *American Antiquity* 57, 3, 1992, p. 109.

12 Hodder, "Interpretative Archaeology and Its Role", in: *American Antiquity* 56, 1991, p. 12.

13 Hodder, *Ibid*. p. 10.

14 *Ibid.* p. 12.

ferent from our assumptions".[15] This allows him to oppose to post-structuralism and post-modernism relativism, since "*we are not interpreting interpretations*".[16] However, his advocacy of objectivity is not easy to hold in the late hermeneutics. Hodder needs some method that allows him to ensure a proper interpretation of the past, beyond his advocacy for self-criticism and the acknowledgment of subordinate voices (such as women or indigenous people) as a subjective requirement. Nonetheless, it is also true that Hodder offers some clues in the way of the method, when he attaches to internal consistency and external experience an important role in evaluating the interpretative hypotheses (as will be seen below).[17]

2.1 Other post-processual answers to Relativism and Subjectivism

Neo-Marxist, structuralist, and post-modern perspectives have given other answers to the problems of subjectivism and relativism. Neo-Marxists give great weight to ideologies to understand the changes in societies of the past (compared with traditional Marxism that gave weight to the economic infrastructure).[18] Neo-Marxist central thesis is that archaeology has a major ideological-political content, but archaeologists must make their interests and beliefs explicit, and be politically responsible for their claims about the past; that is the only possible objectivity. This viewpoint has been important in the emergence of local archaeologies in third world countries insofar as neo-Marxists pay attention to the role of the cultural archaeological past in determining the historic identity of regions.

The object of structuralist archaeology is the structure of the thought which exists in the minds of those who elaborated the artefacts and created the archaeological record (analyses are synchronic). There are constant patterns in human thought in different cultures (dichotomies for example), and the categories observed in one sphere of life will appear in another: culture categories to delimit social relations will also be detected in other different areas, like the delimitations in decorating pottery. In any event, structuralist archaeologists assume that it is possible to access these universal meanings objectively (semiotics) and hence archaeology's interpretations are objectivised.[19]

---

15 Hodder, *Ibid.,* p. 12.

16 *Ibid.,* p. 12 (his italics).

17 In the same, Hodder, "Interpretative Archaeology and Its Role", pp. 7-18, and in: Hodder, "The Post-processual Reaction", in: Hodder, *Theory and Practice in Archaeology*. London: Routledge 1992, pp. 160-168.

18 See, Leone, "Liberation not Replication: 'Archaeology in Annapolis' Analyzed", in: *Journal of the Washington Academy of Sciences* 76, 1986, pp. 97-195. Trigger, "Marxism in Contemporary Western Archaeology", in: *Archaeological Method and Theory* 5, 1993, pp. 159-200. Russell C. Handsman, "Early Capitalism and the Center Village of Canaan, Connecticut: A Study of Transformations and Separations", in: *Artifacts* 9, 1981, pp. 1-2.

19 See Leroi-Gourhan, *Lárt parietal: language de la préhistoir*, *op. cit*., and Preucel, "The Postprocessual Condition", pp. 147-175.

In the 90s, post-modern theses boomed in post-processual archaeology. This movement was inspired by the post-modern philosophy, Critical Theory, post-structuralism and literary criticism, and it took a radical interpretative stance. It opposed not only the processual approach, but also certain schools of post-processual archaeology, such as neo-Marxism or structuralism. Post-modern archaeologists insist on highly relativist and constructivist positions. They consider that any interpretation refers to the outside world; hence the only support for knowledge is a network of interpretations.[20] Archaeologists simply construct their data, and even the facts, from their theories, their cultural present and their subjectivity. Data records should be understood as texts that are interpreted in different ways from different readings made by individuals with different interests, ideologies and beliefs. Interpretations are always presentist, contextual and circulars and all of them have the same right to be sustained, there is no way to establish whether one is more correct or better than any other. Therefore what is being questioned by post-modern archaeologists is the whole edifice of knowledge that characterises the Modernity that must be totally deconstructed.

Finally, several authors have recently opted to inquire about methodological and epistemic criteria that allow the introduction of some degree of objectivity in the field of interpretation avoiding radical relativism, and thus constituting a certain "third way" of the post-processual archaeology.

3. Common ground between processual and post-processual archaeology

Recently, several post processual archaeologists have followed a new way given the sterility of debates between processual, hermeneutics and post-modern archaeologists, according to these authors. This new approach tries to find some common ground between post-processual and processual post-empericist archaeologies, despite of major differences between them. This common ground shapes a third way for archaeology, characterised by a pluralism that could accommodate both, resources from interpretative and post-empiricist approaches. Research into this common ground is making possible that the opposite monolithic positions are starting to lose ground in favour of other more integrating stances. Work has gone in two directions: a) to show that post-processual archaeology can accommodate certain post-empiricist criteria, and b) to show convergence in the archaeological research.

Concerning convergent criteria, VanPool and VanPool argue that given the evolution of the philosophy of science and the post-empiricist reformulation of the

20 See for example: Christopher Tilley, "Interpretation and a Poetics of the Past", in: Tilley (Ed.), *Interpretative Archaeology*. London: Berg 1993, pp. 1-26, and Bintliff, "Postmodernism, Rhetoric and Scholasticism at TAG: The Current State of British Archaeology", pp. 274-278.

criteria of scientificity, post-processual research satisfies many of such criteria;[21] they claim:

> we suggest that much of the discussion of the relationship between processual and post-processual archaeology is based on subtle, yet important, misconceptions (…) we will discuss seven recognized characteristics of science and demonstrate that processualism and postprocessualism both possess most of these characteristics. We therefore suggest that the conflict between the practitioners of processual and post-processual archaeological approaches is largely unnecessary, not because of the social implications of the conflict, but because of their substantive intellectual similarities.[22]

Hutson and Wylie have argued in the same direction;[23] and Fogelin believes that despite the epistemological and theoretical debate that has divided the two archaeological approaches, "those with more interpretive leanings are actively engaging in field work and re-embracing many of the 'scientific' methodologies pioneered by the New Archaeology".[24]

Fogelin, is also representative of the second point indicated above, the convergence in the archaeological research. He considers that "in the meantime, both sides borrow data from one another and continue to rely on the work of archaeologists from the early twentieth century".[25] Thus, despite the different approaches, in practice, there is sufficient proximity in the research and techniques to share the data and to trust the results of the research. In fact, archaeologists share a fairly general agreement regarding the techniques they use, irrespective of which epistemological approach they sustain. On the other hand, it is fairly widely accepted that, despite different points of view, archaeologists have offered a series of powerful explanations of the past. It does not mean that all research was good research, but a set of explanations that are accepted as correct have been established. In fact, as Fogelin notes, many archaeologists consider they are working in a "middle ground" between the processual and the post-processual perspective.[26]

---

21 In their articles, Christine S. VanPool and Todd L. VanPool, "The Scientific Nature of Postprocessualism", in: *American Antiquity* 64, 1999, pp. 33-53; and, T. L. VanPool and C. S. VanPool, "Postprocessualism and the Nature of Science: A Response to Comments by Hutson and Arnold and Wilkens", in: *American Antiquity* 66, 2001, pp. 367-375.

22 C. S. VanPool and T. L. VanPool, "The Scientific Nature of Postprocessualism", in: *American Antiquity* 64, 1, 1999, p. 34.

23 See Scott R. Hutson, "Synergy through Disunity, Science as Social Practice: Comments on VanPool and VanPool", in: *American Antiquity* 66, 2001, pp. 349-369. Wylie, *Thinking from Things, Essays in the Philosophy of Archaeology*, *op. cit.*

24 Lars Fogelin, "Inference to the Best Explanation: A Common and Effective form of Archaeological Reasoning", in: *American Antiquity* 72, 2007, p. 604.

25 *Ibid.*, p. 604.

26 *Ibid*., p. 604.

3.1 Convergences of Explanation and Understanding

Despite the convergences abovementioned, very little analysis has been conducted from this point of view on explanation-understanding debate in archaeology. But, in this area is possible to establish interesting points of convergence.

A first convergence is related to the admission by processual archaeologists that explanation involves auxiliary interpretive hypotheses. They ended up acknowledging that the connection (deductive or inductive) between explanans and explanandum assumed some level of interpretation (and the test of the hypothesis too). Processual archaeologists admitted the importance of the theory laden nature of observation and that all facts implied theoretical-dependent interpretation. The question was pointed out by Binford when he acknowledged that the data talks but they do not talk by themselves of the cultural processes or ways of life unless we ask them the right questions.[27] Archaeologists are not confined to understand, also try to explain facts. But the facts they try to explain have previously been interpreted, either in the context of a paradigm, a theory or a background of knowledge.[28] Interpretation allows understanding the facts of the past which are made intelligible and so explainable. Without understanding the meaning of these facts (to which the archaeologists access from the remains of the present), their explanation is not feasible. The explanation may rely on the insides of past actions, on contextual aspects of material cultures or on constraints to which people were subjected, whose meaning has been established through interpretation. Explanations can be of different kinds, but are possible once the meaning of facts to be explained has been understood. This can be deemed in terms of Weber's proposal, who maintained that social researchers should not be content with interpreting meanings renouncing to explanation. But, in opposition to positivists, he also considered that explanation needs to take in consideration the meaning and sense connections. Thus, in social sciences, understanding is not in opposition to explanation, but rather it constitutes a necessary moment of explanation.

3.2 The Role of the Evaluation

Another important area of convergence between processualism and post-processualism is related to the acceptance by some post-processualist arqueologists of the evaluation of interpretive propositions. Processual archaeologists understand

27 Binford, "Archaeological Perspectives", in: L. R. Binford, and S. R. Binford (Eds.), *New Perspectives in Archaeology*, *op. cit*., p. 13. Hodder later admits that "both processual and hermeneutic approaches accept that every assertion can be understood in relation to a question", Hodder, "Interpretative Archaeology and Its Role", *op. cit.,* p. 12.

28 Meanings can be from individuals or contextual elements; as C. S. VanPool and T. L. VanPool note, "Social meaning can be given to material objects, people, societies, and places through interpretation". C. S. VanPool and T. L. VanPool, "The Scientific Nature of Postprocessualism", *op. cit*., p. 38.

that explanatory propositions had to be empirically tested and post-processual archaeologists maintain that interpretive propositions should be empirically evaluated.

Many post-processual archaeologists admit the empirical reality of entities, and above all, of the archaeological records.[29] As VanPool and VanPool point out: "Most post-processual interpretations meet the requirement of empiricism (…), post-processualists do accept that the past is real and that they can know something about it".[30] Thus, although post-processualists believe that archaeological research is cultural, politically or value-interest basied, many of them accept that the archaeological records limit their interpretations. Archaeological data records constrain interpretations and meanings that can be established, as the same Hodder states, "the real world does constrain what we can say about it".[31] This enable data to play an important role in the evaluation of interpretative hypothesis; interpretation is not a case of anything goes.

Interpretations can be infinite, but archaeologists do not believe that all interpretations are valid. Their evaluation is what makes possible to establish which are considered valid and which are not. Thus, the evaluation criteria are a key resource of archaeological research, and one of the criteria proposed by archaeologists has been *consistency with the archaeological records*,[32] (also the internal coherence and the inference to the best explanation).[33]

The criterion of *consistency with the data* has been understood in different ways. VanPool and VanPool, and Preucel have interpreted it as "adequacy with the inter-subjectively testable data", which entails objectivity.[34] VanPool and VanPool give a strong epistemic meaning to this criterion since they consider that "inter-subjectively testable" has a clear Popperian label, "as Popper (1980: 44)

29 For example, Ericka Engelstad, "Images of Power and Contradiction: Feminist Theory and Potprocessual Archaeology", in: *Antiquity* 65, 1991, pp. 502-514. Leone, "Liberation not Replication: 'Archaeology in Annapolis' Analyzed", *op. cit*., pp. 97-195. Hodder, *Reading the Past: Currents Approaches to Interpretation in Archaeology*, *op. cit.* Hodder, "Interpretative Archaeology and Its Role", *op. cit*., pp. 7-18.

30 C. S. VanPool and T. L. VanPool, "The Scientific Nature of Postprocessualism", *op.cit*., p. 42. Hodder, as has been seen above, holds the same thesis; see Hodder, "Interpretative Archaeology and Its Role", *op. cit.,* p. 12.

31 Hodder, *Reading the Past: Currents Approaches to Interpretation in Archaeology*, *op. cit*., p. 16.

32 Hodder, "Interpretative Archaeology and Its Role", *op. cit*., pp. 7-18. Preucel, "The Postprocessual Condition", pp. 147-175. C. S. VanPool and T. L. VanPool, "The Scientific Nature of Postprocessualism", *op. cit*., pp. 33-53. Wylie, *Thinking from Things, Essays in the Philosophy of Archaeology*, *op. cit.*

33 The first one in Hodder, "Interpretative Archaeology and Its Role", *op. cit*., pp. 7-18, for example. The second one in Fogelin, "Inference to the Best Explanation: A Common and Effective form of Archaeological Reasoning", *op. cit*., pp. 603-625.

34 C. S. VanPool and T. L. VanPool, "The Scientific Nature of Postprocessualism", *op. cit*., pp. 44-45. Preucel, "The Postprocessual Condition", pp. 161-162. (Italics in quote are mine).

states, 'the objectivity of scientific statements lies in the fact that they can be inter-subjectively tested'".[35]

Hodder argues for a weaker recourse to empirical. He states that, "we need to retain from positivist and processual archaeology a guarded 'objectivity' of the material (…) the organized material remains have an independence that *can confront our taken for granted*".[36] The past is objectively organized and it is different from our contexts, and "it is in the experience of this objective and independent difference that we can distinguish among competing hypotheses to see which fits best".[37] Of course, pre-existing beliefs inform the interpretations of the past, but, as Fogelin points out, "the material remains, however, are not amenable to just any interpretation. Some interpretations will be shown wrong through a failure to account for the diversity of evidence that is structured by people in the past".[38]

Arnold and Wilkens are critical of these attempts at convergence. They cast doubt on the analysis of VanPool and VanPool and understand that the term "adequacy with the inter-subjectively testable data" does not mean the same in the scientific method as in the hermeneutic method. They claim that "scientific confirmation, as noted above, is commonly assessed as *a function of the independence between the hypothesis being evaluated, and the methods/ knowledge claims used to render that evaluation*".[39] But interpretive hypotheses are not validated by resorting to external resources; they are validated by resorting to internal meanings to the interpretations themselves.

VanPool and VanPool respond to this criticism by maintaining their original position, and focussing on the question of objectivity, to demonstrate that interpretive hypotheses are tested in independent evidence as much as the descriptive or explanatory hypotheses. They reply that:

> we find this claim startling given the relatively large number of post-processual studies that contradict it (e.g., Hodder 1984, 1992:216-228, 2000:28-30; Marciniak 1999; Pauketat and Emerson 1999; Prestvold 1996; Thomas 1996:98-233). For example, Meskell (1998:229-233) uses spatial analysis, ethno-historical evidence, evidence from other sites, architectural analysis, and contextual analysis to evaluate her contention that certain rooms from Deir el Medina, a New Kingdom settlement in Egypt, were used predominantly by males.[40]

35 C. S. VanPool and T. L. VanPool, "The Scientific Nature of Postprocessualism", *op. cit*., p. 44.

36 Hodder, "Interpretative Archaeology and Its Role", *op. cit*., p. 12.

37 Hodder, *Ibid*., p. 13.

38 Fogelin, "Inference to the Best Explanation: A Common and Effective form of Archaeological Reasoning", p. 613.

39 Philip J. Arnold III and Brian S. Wilkens, "On the Van Pools "Scientific" Postprocessualism", in: *American Antiquity* 66, 2001, p. 363; (their italics).

40 T. L. VanPool and C. S. VanPool, "Postprocessualism and the Nature of Science: A Response to Comments by Hutson and Arnold and Wilkens", *op. cit*., p. 369.

4. Concluding remarks: A third way for explanation

What the arguments in favour of a convergence between processual and post-processual archaeology show is that the differences between the two approaches, which initially seemed abysmal, are reduced insofar that the positions of the two parties become more flexible. As it has been seen, the evolution followed by the former and the moderate considerations of the latter make possible to build bridges at key points between the two perspectives. This does not mean that there do not remain radical postures, fundamentally by the archaeologists of the post-modern thought, or by more scientificist positions, such as those of Arnold and Wilkens who make a radical distinction between scientific method and hermeneutic method.

Furthermore, it is interesting to point out that the debate between Arnold and Wilkens, and VanPool and VanPool about scientific and hermeneutic method has its counterpart in the philosophy of science, although concerning two ways of understanding scientific proof. The discussion about to what extent the scientific proof is internal (given that what is postulated as evidence is determined by theories/hypotheses being tested) or to what extent it is external (insofar evidence is independent from the theories/hypotheses being tested) continues open in philosophy of science.

On the other hand, the arguments in favour of a convergence between processual and post-processual archaeology are interesting for the general philosophical debate between explanation and understanding. What the arguments examined bring to this debate can be synthesised in the following points:

1. Explanation shares with understanding the fact that what is explained or understood means something because it is first interpreted. It is does not mean that the data have no role in the interpretation or in establishing whether an interpretation is more correct or better than any other. The interpretations can be infinite, but archaeologists do not believe that any interpretation is valid (except the post-modern ones). Processual archaeologists understood that interpretive propositions had to be empirically tested and post-processual archaeologists maintain that interpretive propositions should be empirically evaluated.

2. Insofar as explanation in social sciences refers, in one way or another, to actors and their actions (as is widely admitted), the understanding of the actions is prior to their explanation, and constitutes a necessary moment of the explanation. Without understanding the reasons at stake in actions, one cannot proceed to their explanation. This can be understood as a proposal that affects not only to social sciences, but also to sciences of culture, which can include some kinds of explanation for which understanding would be a necessary condition – what would be in accordance with Wright's thesis that understanding is a fundamental condition that comes prior to the teleological explanation of social actions.

3. Explanations that appeal to causes co-exist in social sciences with those that resort to reasons or intentions. It is true, as has been pointed out by Gonzalez,

that the question for the causes and the question for the reasons are different; the causal explanation is formulated to human behaviour, while the question for the reasons is formulated to the actions as activity characterised by its "historicity".[41] But, it is also true that to identify something as a cause involves interpretation and understanding in a relevant sense. This does not mean that both forms of explanation are equal since they respond to distinct objectives. Social sciences include both behaviour and its causes, and actions and their reasons.[42] Some social processes and mechanisms have shown to be interesting to explain certain kinds of behaviour, and therefore social facts, but human activity and its "insides" form part of most explanations in social sciences. On the other hand, as Gonzalez maintains, "combining causal explanations and interpretive perspectives can be seen as an example of the unity and diversity of science: the social sciences can share common ground with the natural sciences and, at the same time, they can also present some differences".[43]

University of La Laguna
Campus de Guajara s/n
38207 La Laguna, Sta Cruz de Tenerife
Spain
agomez@ull.es

41 Wenceslao J. Gonzalez, "Sobre la predicción en Ciencias Sociales: Análisis de la propuesta de Merrilee Salmon", in: *Enrahonar* 37, 2005, p. 193.

42 Amparo Gómez, "Mechanisms, Tendencies and Capacities", in: *Peruvian Journal of Epistemology*, v. 2, forthcoming.

43 Gonzalez, "Sobre la predicción en Ciencias Sociales: Análisis de la propuesta de Merrilee Salmon", p. 188.

Demetris Portides

Idealization in Economics Modeling

Abstract

I argue that understanding idealization as a conceptual act that can be distinguished into three kinds: isolation, stabilization and decomposition is a promising way for making sense of many important characteristics of economic modeling. All three kinds of idealization involve the conceptual act of variable control which amounts to omission of information. I particularly highlight the point that in addition to isolations and stabilizations an implicit (and occasionally explicit) feature of idealization in economics modeling is *decomposition*, i.e. the idea that we set apart within our model descriptions clusters of factors that we assume to influence the behavior of the target system by abstracting from the complex natural (or social) convolution of things in the actual world. These features of idealization are explicated with reference to particular examples of scientific models.

1. Introduction

In the last few decades it has become widely accepted that idealizations enter in a variety of different ways in most aspects of scientific practice.[1] Idealization in scientific models, in particular, has received the most attention primarily because idealizing assumptions enter in the development of models, and since the latter are constructed in order to represent physical systems questions about idealization are linked to questions about scientific representation. In order to make sense of scientific representation questions about the nature of scientific models must be addressed, and in order to illuminate the notion (and functions) of scientific model it is important to address the character of idealization. Despite the widely shared view of the importance of idealization there is, however, no consensus as to how to construe its character and its epistemological implications.

In this essay I shall not address any epistemological questions that arise from the use of idealization in science. Instead I shall focus on the character of idealiza-

1 Amongst others, this is largely due to the work of Ernan Mcmullin, "Galilean Idealisation", in: *Studies in History and Philosophy of Science* 16, 1985, pp. 247-273; Frederick Suppe, *The Semantic Conception of Theories and Scientific Realism*. Urbana: University of Illinois Press 1989; Leszek Nowak, *The Structure of Idealization*. Dordrecht: Reidel Publishing Company 1980; Nancy Cartwright, *Nature's Capacities and their Measurement*. Oxford: Clarendon Press 1989.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_20,

tion, particularly regarding its use in scientific models. More specifically, I will analyze three different kinds of idealization which enter in economic modeling. I do not wish to confine my argument to the discipline of economics, thus I do try to draw attention to the fact that the same kinds of idealization are discernible in other scientific disciplines, e.g. cognitive psychology. Of course, much can be said about scientific models in order to shed light on the notion. The various kinds of models that are encountered in scientific practice, as well as the different ways by which scientific models are constructed and the wide variety of their functions has been the subject of inquiry of a growing number of philosophers in the last few decades.[2] Although there is no agreement among philosophers neither regarding the function of models nor regarding the nature of their relation to their respective target systems,[3] there is no dispute about the fact that models present simplified descriptions of their targets. In other words, that many of the complexities that are present in the actual target systems are most often not included in scientific models.

Simplification in modeling is not, of course, achieved only by idealization. Scientists also simplify by the use of mathematical approximations and possibly by other means. I shall not concern myself with other kinds of simplifications here and neither with how these blend together with idealization in the construction of scientific models. My main concern in this paper is to discern the kinds of idealizations involved in scientific modeling with particular emphasis on economic models. I defend the claim that the categorization, I suggest, can improve our understanding of modeling in economics as well as in other disciplines. A further goal of this paper is to highlight a kind of idealization that has received little attention in the literature, which I call *decomposition*.

---

2 See for instance, Ronald Giere, *Explaining Science: A Cognitive Approach*. Chicago: The University of Chicago Press 1988; Margaret Morrison, "Models as Autonomous Agents", in: Mary Morgan and Margaret Morrison (Eds.), *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press 1999, pp. 38-65; Cartwright, *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press 1999; and Newton da Costa and Steven French, *Science and Partial Truth.* Oxford: Oxford University Press 2003.

3 The more recent attempts to explore the functions of models also allow, for example, older philosophical debates to be reborn albeit within a new framework and a new language. One such instance is the debate on the methodological character of economics, and more generally the social sciences, on the significance of prediction as opposed to understanding (see Wenceslao J. Gonzalez, "From *Erklären-Verstehen* to *Prediction and Understanding*: The Methodological Framework of Economics", in: Matti Sintonen, Petri Ylikoski and Karlo Miller (Eds.), *Realism in Action: Essays in the Philosophy of Social Sciences*. Dordrecht: Kluwer 2003, pp. 33-50).

2. Idealizing assumptions in modeling

Uskali Mäki makes use of von Thünen's classic economic model of the isolated state in order to explore the notion of idealization. Von Thünen invites his reader to:

> Imagine a very large town, at the center of a fertile plain which is crossed by no navigable river or canal. Throughout the plain the soil is capable of cultivation and of the same fertility. Far from the town, the plain turns into an uncultivated wilderness which cuts off all communication between this State and the outside world. There are no other towns on the plain.[4]

The above quote is suggestive of the characteristics of von Thünen's economic model. Uskali Mäki goes on to spell out the assumptions that lie beneath it. Here is Mäki's complete list of assumptions:

> 1) The area is a perfect plain: there are no mountains or valleys. 2) The plain is crossed by no navigable river or canal. 3) The soil in the area is throughout capable of cultivation. 4) The soil in the area is homogeneous in fertility. 5) The climate is uniform across the plain. 6) All communication between the area and the outside world is cut off by an uncultivated wilderness. 7) At the center of the plain there is a town with no spatial dimensions. 8) There are no other towns in the area. 9) All industrial activity takes place in the town. 10) All markets and hence all interactions between the producers are located in the town. 11) The interaction between producers is restricted to the selling and buying of final products: there are no intermediate products and no non-market relationships between producers. 12) Transportation costs are directly proportional to distance and to the weight and perishability of the good. 13) All prices and transportation costs are fixed. 14) Production costs are constant over space. 15) The agents are rational maximizers of their revenues. 16) The agents possess complete relevant information.[5]

Mäki asks what functions these assumptions serve, since they are obviously false if they are about a real economy. His answer reveals how he conceives idealization and its function in scientific modeling: "… the function of such falsehoods is isolation by idealization … Idealizing assumptions 1-16 serve the function of neutralizing a number of causally relevant factors by eliminating them or their efficacy".[6] Thus what modelers do, according to Mäki, is completely omit some factors relevant to the behavior of the target system or omit some characteristics of retained factors that are influential in the target's behavior. The final result of such a conceptual act he calls *isolation*. Mäki is not the only philosopher to understand the character of idealization along these lines.

---

4 From Uskali Mäki, "Models and the Locus of their Truth", in: *Synthese* 180, 2011, pp. 47-63, p. 50.

5 *Ibid.* p. 50.

6 *Ibid.* p. 51.

Ernan Mcmullin, for example, claims that, "every theoretical model idealizes, simplifies to some extent, the actual structure of the explanandum object(s). It leaves out of account features deemed not to be relevant to the explanatory task at hand".[7] He goes on to distinguish between two different ways this is done. He labels the case when the features simplified or omitted are known to be relevant to the kind of explanation aimed by the model, *formal idealization*; and the case when the unspecified features are considered irrelevant to the inquiry at hand, *material idealization*. Mcmullin's distinction between formal and material idealization, as well as what he calls *subjunctive idealization*, concerns the levels of language at which idealizations enter. In other words, material idealizations do lie beneath the construction of a model but are usually dictated by the theory for which the model is an application, i.e. they are necessary conditions for theory application within a particular domain that *inter alia* provide the theoretical underpinnings of the model. On the other hand, formal idealizations – and the same goes for subjunctive idealizations – are explicit or implicit assumptions used to set-up the model *per se*, i.e. they are contingent features of the theory application the model is meant for or of the problem situation the model aims to provide a solution for. In all cases (and in particular in formal idealization, which is the kind of idealization that is primarily operative at the level of modeling) 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, e.g. omitting frictional effects altogether from the description of a physical system. What he calls simplification is the kind of idealization where relevant features that are retained in the model description are represented in the model equation(s) in a more simplified way than the way they are perceived in the target system, e.g. representing a body in a physical system as if it has infinitesimally small spatial extension. Both Mcmullin and Mäki blend simplification in this sense and omission into their generic notion of idealization.[8]

The minor differences in their views, as well as their terminological differences, can be attributed to the different perspectives they have. Mcmullin conceives idealization as an act that does not change the reference of the language of the model (thus he views it as a form of simplification) because he conceives the model as relating more or less directly to its target. Whereas, Mäki sees idealization as shifting the reference of the language of the model from the target to an 'imaginary' situation which is indirectly related to the actual target (thus he calls

7 Mcmullin, *Ibid*. p. 258.

8 In fact, many authors blend the two notions. For instance, Nowak, *Ibid.*, blends them into his notion of 'idealization'. Similarly, Morrison, "Models, Pragmatics and Heuristics", in: *Dialektik* 1, 1997, pp. 13-26, blends them into her notion of 'computational idealization;' and Steven French and James Ladyman, "Semantic Perspective on Idealisation in Quantum Mechanics", in: Niall Shanks (Ed.), *Idealisation IX: Idealisation in Contemporary Physics, Poznan Studies*. Vol. 63, Amsterdam: Rodopi 1998, pp. 51-73, also blend them into their notion of 'idealization'.

the result of idealization, 'isolation'). These differences, although interesting, are of no importance to my concerns in this paper. What is of importance is the fact that both authors conceive idealization as the act of omitting features of the target system from the model description. Furthermore, that the act of omission can be distinguished into two sorts: complete omission of a relevant factor to the target's behavior or omission of characteristics that lead to simplification of a relevant factor that is retained in the model description.[9]

Although I do not dispute the general way by which Mcmullin and Mäki conceive the methodological aspects of idealization, I do think that the picture they present does not sufficiently capture important elements of modeling. There are two reasons for this lack of adequacy. The first concerns what Mcmullin calls 'simplification' of features or what Mäki calls elimination of the 'efficacy' of a factor. To clarify these notions for the purpose of understanding the character of idealization I think one should address "what is involved in simplifying a feature" or "what is involved in eliminating the efficacy of a factor". However, even if adequate answers are given to these questions we are still not led to an adequate conception of idealization as employed in modeling by relying solely on the above two sorts of omission. I shall argue that adequacy is achieved when we augment our conception of idealization with the idea of decomposition. I shall try to clarify these points with reference to the *isolated state* model in the next section.

3. Three kinds of Idealization: *Isolation, Stabilization, Decomposition*

A closer examination of the assumptions underlying von Thünen's model reveals that the ways by which they differ allow various ways by which one could categorize them, depending on one's perspective. Mäki's idea that the underlying idealizing assumptions all serve the function of isolation, i.e. they enable the construction of a description that refers to a situation in which some of the factors and characteristics of the target system are conceptually screened-off from the rest, may be found useful if one is interested in questions regarding what the model refers to and how that relates to the actual target. However, if one is interested in questions concerning the cognitive act involved in the introduction of idealiza-

9 Not all philosophers agree with this idea. For example Cartwright, *Nature's Capacities and their Measurement*, claims that two distinct thought processes are involved; that of idealization, which she conceives as the act of distortion of the target of a model, and that of abstraction, which she conceives as the act of omitting causally relevant factors from the model description. Similarly, Suppe, *Ibid.*, makes the same distinction on roughly the same grounds. Although I shall not argue for this, I side myself with Mcmullin and Mäki on this issue and understand idealization as the conceptual act of abstracting from the complexities of the target system, or purposefully eliminating factors altogether or some of their features from the model description.

tions for the construction of models then the notion of isolation (in Mäki's sense) will not suffice, since itself is the result of the act of conceptually omitting factors, i.e. controlling the variability of parameters.[10] For example, assumption (6) is a straight-forward kind of omission that screens-off all communication between the area and the outside world. Entirely omitting influences from the outside world to the economy the model aims to describe does not prohibit talk about economy. Assumption (7), however omits the spatial dimensions of the town. Any model must talk about something, in this particular case it must talk about a location in which all economic transactions take place, but the dimensions of the town can be omitted without making talk about economic relations unsound. Furthermore, different assumptions concern different processes of the economy. For instance, assumptions (3), (4) and (5) concern production, whereas assumptions (1), (2), (12) and (13) concern transportation, and assumptions (10), (15) and (16) concern exchange. By noting these differences, we could usefully categorize the idealizing assumptions, from the perspective of the thought process, into three kinds: isolation, stabilization, and decomposition.

Isolation is the act of abstracting from some relevant factors that are found in the target system or conceptually omitting entire characteristics of the target system, i.e. setting the value of variable parameters to zero. In von Thünen's model several assumptions involve an isolation component. For example, assumption (3) omits all possible obstacles of cultivation, e.g. large rocks. Similarly, assumption (2) omits all rivers and canals and assumption (6) omits all kinds of communication with the outside world, which amounts to omitting the influence on any of the component parts of the model from an outside world. Notice that, although I borrow the term from Mäki, what I call here 'isolation' is the *act* of conceptually screening-off a certain situation from some factors that are present in the actual situation the model is meant to represent, and not the final conceptual construct as Mäki uses the term. Some authors have referred to this act as 'abstraction' but I think this will not do because abstractive analyses are much more general and seem to be involved in all three kinds of idealization.

Generally, isolation consists in the omission of existence claims about the influence of factors – or of entire characteristics of target systems – on the investigated behavior of a phenomenon, i.e. in isolation we omit the information present in the claim that "there exists a factor x that influences the investigated behavior y". Cognitive psychology is another of many scientific areas in which isolation is vividly present. For example, in Piaget's work the development of human cognitive capacities and abilities is treated as a series of distinct stages from infancy to adolescence, i.e. from sensor motor stage to the more complex formal thinking and knowledge acquisition. Such treatment relies on assumptions that lead to an ideal-

10 This is not a claim that there is one perspective of analyzing idealization which is of utmost significance, but I do think that from the perspective of the cognitive act (or thought process) behind idealizing assumptions we can learn something of interest and value to idealization and to scientific modeling.

ized model for cognitive development since cultural, historical, and geographical influential factors are entirely screened-off.[11]

Stabilization is the act of conceptually omitting some of the characteristics (constitutive parts) of a factor involved in the behavior of the target system while retaining the factor itself in the model description, albeit without the features omitted.[12] This kind of idealization occurs when we abstract from the significance of the naturally occurring magnitude of a characteristic or when we abstract from the variability of a characteristic. In von Thünen's model several assumptions involve a stabilization component. For instance, assumption (1) involves the claim that the area is a perfect plain, in other words the variability of the angle of inclination of the area is omitted, i.e. the area is assumed to have constant inclination. In addition to this omission, the assumption involves a synthesis of information, namely a particular value is given to the angle of inclination, that of zero. Assumption (4) involves the claim that the soil's fertility is homogeneous, i.e. all "impurities" that would otherwise disturb the fertility of the soil are omitted (or more generally, it abstracts away from the variability of fertility). Such assumptions do not eliminate the features of 'being an area of cultivation' or 'having fertility' altogether, thus allowing the model to talk about them albeit not with all their actual characteristics. Similarly, assumptions (15) and (16) involve the claims that the agents are rational maximizers of their revenues and that they possess complete relevant information, by abstracting away from the fact that rationality (in the economic sense) and possession of relevant information are variable characteristics and furthermore attributing to these aspects of agents extremum values. These assumptions also do not altogether eliminate the features of an agent having 'cognitive ability' or 'informational capacity', but they alter some of the actual characteristics of these features.

Some authors view such idealizations as often involving a distortion of the actual characteristics of a factor.[13] Although this may be the case if such idealizing assumptions are examined from the point of view of the conceptual construct, if one examines them from the perspective of the conceptual act that gives rise to the final construct then the proper question to ask is: "What is common to such acts of shaping conceptual constructs for the purposes at hand?" That which is common to such assumptions as the ones described is that they involve the omission of the

11 Jean Piaget, *The Origins of Intelligence in Children*. International Universities Press: New York 1952.

12 I borrow the term 'stabilization' from Renata Zielinska, "A Contribution to the Characteristics of Abstraction", in: Jerzy Brzezinski, Francesco Cogniglione, Theo Kuipers and Leszek Nowak (Eds.) *Idealisation II: Forms and Applications, Poznan Studies*. Vol. 17, Amsterdam: Rodopi 1990, pp. 9-22, in which it is used to express a more or less similar idea.

13 E.g. Cartwright, *Nature's Capacities and their Measurement*, *op. cit.*; Michael Weisberg, "Three Kinds of Idealization", in: *The Journal of Philosophy* CIV, 2007, pp. 639-659.

information that a quantity is variable synthesized with the information that the quantity takes on an extremum value (the latter frequently conflicts with the information that there exists a natural upper or lower bound to the parameter). In other words, there is a limit to how small the angle of inclination of a piece of land of unspecified dimensions can be; never mind the fact that the angle of inclination of any actual piece of land found on an almost spherical surface cannot be a constant quantity. Similarly, there is a limit to the actual fertility of the soil being homogeneous, as the soil structure and the soil ingredients vary (sometimes significantly) from place to place. Finally, not all agents have the same cognitive ability and not all are uniformly informed; furthermore, there is a limit to the cognitive ability of an agent, and also there is a limit to how informed an agent can be. Removing such information is equivalent to abstracting from the variability of the factor and omitting the information that there exists a limit to how a factor is or manifests itself in actual circumstances in relation to one or more of its characteristics and at the same time attributing to the factor a particular value. An equivalent way to express the kind of idealization involved in stabilizations is this: we omit the information present in the claim that "there exists a variable characteristic x of a factor y the values of which are naturally bounded and influence the investigated behavior z" and attribute to the characteristic a particular value (often an extremum).

Stabilization is also present in modeling in most scientific disciplines. For instance, in cognitive psychology and in particular in Piaget's theory of cognitive development, the development is assumed to occur as if it follows an invariable pathway through which children develop their thinking repertoire. There is enough evidence, however, to know that in reality much of cognitive development follows a far from smooth and linear course. In other words, the course of cognitive development is variable, and the assumptions upon which Piaget's model relies abstract from the variability and consider cognitive development as if it follows a linear course.

So far I have analyzed idealization as isolation, which is present in both Mäki's and Mcmullin's analyses; and idealization as stabilization, which – as I suggest – is a particular way to understand Mäki's 'elimination of the efficacy of a factor' and Mcmullin's 'simplification of a factor'. However, the primary difference of the conception of idealization I suggest lies in the third kind of idealization, which I think is most often present in scientific modeling; and its presence is the reason why I find both Mäki's and Mcmullin's analyses insufficient for capturing the full extent of idealization.

Decomposition is the conceptual act of setting apart factors, clusters of factors, processes, or mechanisms in our model descriptions; a rough way to put it, is that decomposition is the conceptual act of abstracting from interconnection and interaction. Decomposition is rarely an explicit feature of the model or of the idealizing assumptions that underlie the model. Most of the times they are implicitly present in the underlying assumptions. They are implicit features of the model when we separate the effects of a common cause, or when we separate the

causes of a particular effect, but they are also implicit features of the model when we construct a manifold of – frequently incompatible – models in order to investigate different properties of the target by the use of different models. For instance, when a body performs translational motion in air and we model it by removing air from our description, in order to improve our representation of the target we reintroduce the effects of air, e.g. impedance of the motion due to air resistance, effects of the buoyancy in air on the motion of the body, rotational motion due to inhomogeneous air disturbances and so on. However, it is not possible to separate these effects in an experimental set-up, when air is present then all of the above three effects, as well as others, are present. In other words, the decomposition is conceptual and we do it for simplification purposes.[14]

In von Thünen's model decomposition is not explicitly stated but it is tacit in the underlying model assumptions. The model treats the economy of the isolated state as if the following processes are decomposed: production, transportation and distribution, exchange and consumption. To get a clear indication of the decomposition of these processes in the model we must read behind the lines of each of the idealizing assumptions. For instance, assumption (2) that omits the presence of a navigable river or canal, in addition to involving an isolation, concerns the process of transportation and treats it as if it is independent of other processes. Similarly, assumption (3) that assumes uniform cultivation of the soil, in addition to involving a stabilization, concerns only the process of production and treats it as if it is independent of other processes. Similarly, assumption (16) that ascribes to agents complete relevant information, in addition to involving a stabilization, concerns the process of exchange and consumption and treats it as if it is independent from other processes. These are just examples to emphasize the tacit presence of decomposition in the assumptions of the model. I do not mean to suggest that each and every assumption of the model concerns only one of these processes and sets it apart from the rest. Some of the assumptions of a model may concern only one process others may concern more than one.

My point is that although the various assumptions used in setting up the model involve explicitly stated stabilizations and isolations, they also tacitly involve decomposition. Von Thünen's model of the 'isolated state' does not just isolate the economy of a town from outside influences but it also decomposes the processes of production, transportation and exchange within the economy it describes and treats them as if they are parallel and independent processes. The presence of decomposition can also be discerned by studying the assumptions and their consequences in clusters. For example, assumptions (13) and (14) which state that prices and transportation costs are fixed and production costs are constant over space imply *inter alia* that no technological change takes place, thus no change

14 Mcmullin, *Ibid.*, comes close to the idea of idealization as decomposition in what he dubs 'subjunctive' idealization, by which he means conceptually setting apart causal lines. Decomposition, however, is much more general; and subjunctive idealization seems to me to be one of its particular modes.

in production or transportation. Hence, the model implies no interaction between the three processes, i.e. production does not affect transportation and thus does not influence exchange. We could therefore conclude that there is a hidden assumption that the three processes operate parallel to each other and in tandem to produce the economic behavior of the isolated state.

Idealization in its decomposition form is also present (in a rather explicit way) in cognitive psychology. For example, Sternberg's theory of intelligence relies on the conception of intelligence as having a triarchic structure based on the following three components: analytical, creative, and practical.[15] Analytical intelligence is componential, that is it allows human cognition to "break down" reality into different parts or components. Creative intelligence involves insight, intuition, and generally a 'divergent' form of thinking. Sternberg's third type of intelligence may have both analytical and creative elements, but more importantly it is contextual; practical intelligence leads to solutions to everyday problems. Decomposing the structure of intelligence as Sternberg does involves a conceptual act, the outcome of which is reflected in the constitutive elements of the specific model of intelligence.

Decomposition is present in modeling more often than not. By decomposing the processes we consider responsible for the observed behavior we are performing an idealization. Generally, decomposition consists in the omission of information present in the claim that "there exists a single convolution of factors x that is responsible for the observed behavior y," which in practice leads modelers to break down x into component parts $x_1, \ldots, x_n$ and to treat the latter as if they consist of disconnected modules or clusters of factors (or processes or mechanisms) acting independently of each other and in tandem to produce the investigated behavior. By decomposition we simplify this complex convolution by omitting information about naturally occurring interconnections and interactions and construct descriptions that purport to represent the complexity as the outcome of independent clusters of factors acting in tandem. So decomposition also involves analysis and synthesis. Clusters of factors are set apart and then placed together to construct an amalgam of components that purportedly produces the behavior of the target. Of course, interaction is often reintroduced into models as an addendum that aims to correct their predictions.[16] In such cases the interaction term of models involves the interaction between the component parts of the model which are not necessarily component parts of the natural or social world, but often are the conceptual creations of modelers that aim for simplification.[17]

15 See Robert Sternberg, *Beyond IQ: A Triarchic Theory of Intelligence.* Cambridge: Cambridge University Press 1985.

16 In Quantum Mechanics this reintroduction very often falls within the realm of perturbation theory.

17 As I mentioned earlier, decomposition is usually an implicit feature of the idealizing assumptions of a model. Only rarely is decomposition a relatively explicit feature of a model. Most examples I know of such kind are to be found in physics and in particular

4. Conclusion

I have argued for understanding idealization as a conceptual act that can be distinguished into three kinds: isolation, stabilization and decomposition. All three kinds of idealization involve the conceptual act of variable control which amounts to omission of information. I have also argued that, in addition to isolations and stabilizations, an implicit feature of idealization in economics modeling is *decomposition*, the idea that we set apart within our model descriptions clusters of factors that we assume to influence the behavior of the target system by abstracting from the complex natural (or social) convolution of things in the actual world.

Recognizing the presence of decomposition in idealizing assumptions involved in scientific modeling is itself an issue of interest. Moreover, recognizing that decomposition is a kind of idealization that is present (whether explicitly or implicitly) not only in physics but also in economics, and in other sciences, leads us to acknowledge that not only idealization in its isolation or stabilization form is common to scientific modeling in general, but also decomposition. Independent of the scientific domain in which decomposition is employed its mere presence gives rise to problems regarding the truth of the propositions that may be extracted from scientific models. This, however, is a separate issue which goes beyond the purpose of this work.

Department of Classics and Philosophy
University of Cyprus
P.O. BOX 20537
1678, Nicosia
Cyprus
portides@ucy.ac.cy

quantum mechanical modeling. I have explored one case of explicit decomposition in Demetris Portides, "Why the Model-Theoretic View of Theories Does Not Adequately Depict the Methodology of Theory Application", in: Mauricio Suarez, Mauro Dorato, and Miklos Redei (Eds.), *EPSA Epistemology and Methodology of Science, Volume 1*. Dordrecht: Springer 2009, pp. 211-220.

Ilkka Niiniluoto

On the Philosophy of Applied Social Sciences

Abstract

The distinction between basic and applied research, widely used for the purposes of science policy, is notoriously vague and ambiguous. In earlier papers, I have argued that there is nevertheless a viable and systematic way of separating these two types of research.[1] An important form of applied research includes design sciences or "sciences of the artificial" in the sense of Herbert Simon.[2] Applied social sciences, which pursue knowledge with the purpose of influencing social behavior and social institutions into a desired direction, can be counted as important examples of such design sciences.

1. Research and development

The OECD office introduced in 1966 definitions which have ever since been widely used within science policy. *Research* (R) is defined as "the pursuit of new knowledge", and *development* (D) is the use of results of research "to develop new products, methods, and means of production".

Historically the division of R and D can traced back to Aristotle's distinction between knowledge (Gr. *episteme*, Lat. *scientia*) and productive arts (Gr. *techne*). For a scientific realist, the R&D divide is essentially the same as the distinction between science and technology: *science* is systematic and critical knowledge-seeking by research, and *technology* is the design and use of material and social artifacts, the art and skill of this activity, and its products.[3] In these terms, development is the same as science-based technology.

For pragmatists and instrumentalists, the situation is different: science is seen as a problem-solving activity, which uses Operations Research (OR) as its typical method. In my view, this blurring of R and D can be avoided if we make a proper distinction between cognitive and practical problems: solution of the former are

1 See Ilkka Niiniluoto, "The Aim and Structure of Applied Research", in: *Erkenntnis* 38, 1993, pp. 1-21, and Ilkka Niiniluoto, "Approximation in Applied Science", in: Martti Kuokkanen (Ed.), *Idealization VII: Structuralism, Idealization and Approximation*. Amsterdam: Rodopi 1994, pp. 127-139.

2 See Herbert Simon, *The Sciences of the Artificial*. Cambridge (Mass.): The MIT Press, 1969. (2nd ed. 1981).

3 See Ilkka Niiniluoto, *Is Science Progressive?* Dordrecht: D. Reidel, 1984, Ch. 12.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_21, © Springer Science+Business Media Dordrecht 2013

new knowledge claims, and those of the latter human decisions to act in a certain particular situation.[4]

R&D is today associated with "national innovation systems". In economics, following Schumpeter, *innovation* means the development of technical discoveries into profitable market products or commodities. A recent definition used in Finland states that innovation is "an exploited competence-based competitive asset".[5] In this sense, innovation is a part of development (D) which is usually processed in industrial laboratories. In Finland, mainly due Nokia's investments, privately funded industrial development covers about 70% of R&D.

The so called "social engineering" and "culture industry" are also parts of the innovation system. Cultural and social sciences may produce as their outcomes cultural and social innovations, if their results are developed in the public or private sector. Examples include democracy, public school, Finnish comprehensive school, social security, the Nordic welfare state, child day-care, maternity pack and clinics, and social media.[6]

2. Basic vs. applied research

The OECD manual made a further distinction between two types of research: *basic* or *fundamental* research is systematic search of knowledge "without the aim of specific practical application", and *applied* research "pursuit of knowledge with the aim of obtaining a specific goal". The former is often characterized as "curiosity-driven" or "blue skies research", the latter as "goal-directed" or "mission-oriented" research.

This distinction is not part of the classical legacy of science, since Aristotle's famous division of theoretical and practical philosophy is quite different: practical sciences, which are concerned with the goals of good human life, include ethics, economy, and politics.

The OECD definitions serve to separate applied research from development (technology) and applications of science (innovation), but they are stated in vague and ambiguous terms. Even the "purest" research is in some cognitive sense goal-directed, and even mission-orientation involves some element of curiosity. There may differences in the speed of utilization of research results, but otherwise the term "aim" could refer to more or less accidental personal motives and knowledge

4 See Ilkka Niiniluoto, "The Foundations of Statistics: Inference vs. Decision", in: Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Michael Stöltzner, and Marcel Weber (Eds.), *Probabilities, Laws, and Structures*. Dordrecht: Springer 2012, pp. 29-41.

5 See *Evaluation of the Finnish National Innovation System: Policy Report*. Helsinki: The Ministry of Education and The Ministry of Employment and Economy, 2009, p. 23.

6 See Ilkka Taipale (Ed.), *100 Social Innovations from Finland*. Helsinki: Baltic Sea Centre Foundation, 2006.

of individual scientists, or to the goals of research sites (university vs. research institute) or funding institutions. It is thus no wonder that this division has been heavily criticized. For this reason, it is worth while to ask whether the proposed pragmatic division could be replaced by a systematic distinction.[7]

3. Utilities

One approach might be based on utilities, understood not as variable personal or institutional motives but as objective standards for assessing quality or success. The science – technology division is reflected in the separation of *epistemic utilities* (like truth, information, truthlikeness, confirmation, understanding, explanatory power, predictive power, simplicity) and *practical utilities* (effectivity of a tool in relation to its intended use, economic cost-benefit efficiency, ergonomical, ecological, esthetic, ethical, and social criteria). The former are relevant for the knowledge claims in science, the latter are principles to be used in technology assessment.[8]

Applied research can be assessed both by epistemic utilities (it pursues knowledge by usually applying the result of basic research) and practical utilities (its knowledge has instrumental relevance for some human activity). This can be seen in typical examples of natural and social applied sciences: engineering sciences, agricultural and forestry sciences, biotechnology, nanotechnology, clinical medicine, public health, pharmacology, nursing science, didactics, pedagogy, applied psychology, social policy studies, social work, political science, business economics, communication studies, development studies, urban research, library science, peace research, military research, and futures studies.

At the same time it is again important to emphasize that all science is not applied. In his classification of sciences Jürgen Habermas suggests that natural science is governed by the "technical interest" of controlling nature.[9] This idea was indeed the key to Francis Bacon's 1620 vision that knowledge of causal laws allows us to control nature and "to subdue the necessities and miseries of human life", but it was in fact only realized at the end the 19th century by new engineering and agricultural sciences. To elevate this model of applied science to a principle of all natural science is to assume the instrumentalist view of science which ignores the theoretical or epistemic interest of scholarly activities.[10]

---

7 See Niiniluoto, "The Aim and Structure of Applied Research", *loc. cit*.

8 See Isaac Levi, *Gambling With Truth: An Essay on Induction and the Aims of Science*. New York: Alfred A. Knopf 1967, and Paul Durbin and Friedrich Rapp (Eds.), *Philosophy and Technology*. Dordrecht: D. Reidel 1983.

9 See Jürgen Habermas, *Knowledge and Human Interests*. Boston: Beacon Press 1972.

10 Cf. Niiniluoto, *Is Science Progressive?*, *op. cit*., p. 221.

4. Descriptive Science vs. Design Science

Another approach is based on the logical structure of the knowledge claims in basic and applied research. Fundamental research is *descriptive science* in the sense that it describes reality (nature, mind, and society) by establishing singular facts about the past and the present and general laws (deterministic and probabilistic) about natural and social systems.[11] Typical causal laws of the form

(1) X causes A in situation B

can be used for the purposes of *explanation* (A has occurred in B because X) and *prediction* (A will occur in B after X).

Examples of descriptive sciences include physics, chemistry, geology, biology, ecology, medicine, history, ethnology, anthropology, psychology, legal dogmatics, sociology, and social psychology.

Predictive sciences, which develop and use methods for predicting and forecasting future events and phenomena, include predictive astronomy, meteorology, social statistics, econometrics, and futurology. They are descriptive sciences which traditionally have been regarded as examples of applied science.

Herbert Simon in 1969 was perhaps the first who called attention to another type of applied sciences: the "sciences of the artificial" are not concerned with how things *are*, but "how things *ought to be* in order to attain goals, and to function".[12] They can be called *design sciences*, in the broad sense that design is concerned with shaping and planning artificial human-made systems (e.g., engineering design, environmental and social planning). As attempts to seek knowledge about design activities, design sciences should not to be confused with science-based design itself. In the same way we have distinguished above science from technology and practical problem-solving.

Thus, my proposal is to define descriptive science so that it includes basic research and predictive science, and applied research so that it includes predictive science and design science.

Design sciences usually have instrumental relevance to some professional practices and arts. For example, the profession of nurses practices nursing and the related art of caring the patients, and their activity can be studied and hopefully improved by nursing science. Similarly, we have the combinations politician/administrator – politics – political science, merchant – trade – business economics, soldier – warfare – strategy – military science, and librarian – library work – library science.

A profession Z, as a human or social activity, can of course be studied from many perspectives, among them the history of Z, the psychology of Z, the sociol-

11 This realist view is opposed to social constructivism which claims that scientific facts are artificial productions of scientific investigations. Cf. Ilkka Niiniluoto, *Critical Scientific Realism*. Oxford: Oxford University Press, 1999, Ch. 9.

12 See Simon, *op. cit*., p. 7.

ogy of Z, the economics of Z, and the ethics of Z. Some of these perspectives, which are usually included in the professional educational programs for Z, belong to fundamental basic sciences. But design science can be viewed as the practical kernel of Z-studies which has the goal of improving the practice or art Z.

These observations also explain the typical historical emergence of design sciences by the "scientification" of Aristotelian productive arts.[13] First the practical skills are based on cumulative everyday experience and trial-and-error, then they are expressed by rules of thumb which are further developed into guide books. The next step is the scientific study of the rules by testing their efficacy and function with experiments.

An example is provided by *evidence-based medicine* (EBM): a medical doctor applies conditional commands or rules of the form

(2) If patient has symptoms S, use treatment X!

Such rules as such are not true or false, but we can gather clinical evidence for their validity by testing whether X cures or heals the disease with symptoms S without side effects. The implicit value premise of (2) that medicine wishes to maintain and improve health is presupposed. Basically the same model of *evidence-based practice* (EBP) can be applied in nursing science.[14]

A similar account can be given for *evidence-based policies* in society. Such principles formulate policy recommendations relative to evidence justified by statistical and social scientific research. When this kind of up-to-date critically evaluated scientific knowledge is disseminated to decision-makers, and the values used decisions are democratically negotiated, legitimate improvements can accomplished in environment, population, housing, education, health, economy, work, and services.

5. Technical norms

Already Simon hinted that design sciences are a special kind of normative science which give us justified knowledge about means – ends relationships. In my view, this idea can be expressed by formulating the knowledge claims of applied design sciences by conditional recommendations of the form

(3) If you want A, and believe that you are in situation B, then you ought to do X.

13 See Ilkka Niiniluoto, "The Emergence of Scientific Specialties: Six Models", in: W. E. Herfel, W. Krajewski, I. Niiniluoto, and R. Wojcicki (Eds.), *Theories and Models of Scientific Processes*. Amsterdam: Rodopi 1995, pp. 127-139.

14 See Ilkka Niiniluoto, "Vårdvetenkapen – vetenskapsteoretiska anmärkningar", in: Kristian Klockars and Lars Lundsten (Eds.), *Begrepp om hälsa*. Stockholm: Liber, pp. 103-114; Sam Porter and Peter O'Halloran, "The Use of and Limitation of Realistic Evaluation as a Tool for Evidence-Based Practice: A Critical Realist Perspective", *Nursing Inquiry* 19, 1, 2012, pp. 18-28.

G. H. von Wright calls such statements *technical norms*.[15] Even though unconditional recommendations of the form "You ought to do X!" or "Given B, you ought to do X!" lack truth values, technical norms are true or false, depending on whether X causes A in situation B. As statements with a truth value, they can be results of scientific research. The technical norm (3) can be justified from above (by deriving it from a basic theory or law of the form (1)) or from below (by supporting the generalization (1) by empirical or experimental experience).[16] It is important that such justification can be *value-neutral* in the sense that commitment of the researcher to the value A is not needed. Still, the conditional norm (3) is *value-laden* in the sense that it essentially involves a value premise as its antecedent. In von Wright's terminology, a person, who accepts the value of A and believes to be in situation B, has a "technical ought" by the norm (3).

The formulation of technical norms involves the language of actions and oughts. Thus, they presuppose the idea of agent causality: X is a factor or variable which can be manipulated by us. Design research makes sense only with respect to artificial systems where human intervention is possible.[17]

Von Wright was primarily concerned with the case where X is a necessary cause A. Variants of (3) can be given in cases where X is a sufficient cause of A (so that it is rational to do X) or X is a probabilistic cause A (so that it is profitable to do X). In the most general case, the end A is expressed by a utility function, the situation B by an epistemic probability distribution over states of nature, and the recommendation of doing X is relative to the conception of rationality (such as minimax or expected utility).

A special issue for applied social sciences is the question whether there are laws about society which can serve as basis of social technical norms. The existence of such laws is often denied by noting that acting against prevailing social trends is always possible at least in principle, so that they are at best ideological or social constructions. However, for the purposes of applied social science, deterministic and permanent "iron laws" are not needed, but temporary statistical regularities in human behavior may be enough. Still, the manipulability condition presupposes that such regularities are just not accidental constant conjunctions but are based upon propensities or some sort of generative powers of causal mechanisms.

15 See Georg Henrik von Wright, *Norm and Action*. London: Routledge and Kegan Paul 1963. For a treatment of so-called technological imperatives as technical norms with a hidden value premise, see Ilkka Niiniluoto, "Should Technological Imperatives be Obeyed?", in: *International Studies in the Philosophy of Science* 4, 2, 1990, pp. 181-189.

16 Illustrations of both of these derivations in the case of ballistics are given in Niiniluoto, "Approximation in Applied Science", *loc. cit*.

17 Theo Kuipers formulates design laws as causal regularities of the form "Functional property A in situation B can be achieved by imposing structural property X", where the term "imposing" involves agent causality. See Theo Kuipers, "Philosophy of Design Research", forthcoming in *EPSA 2011*.

6. Values in applied social sciences

The traditional ideal of value-free science has often been challenged in the context of the social sciences, where the researchers have social positions and political interests. Even though social scientists can empirically study the valuations of human beings in various cultures, it is not legitimate to appeal to one's own values as grounds for accepting or rejecting scientific hypotheses. For descriptive sciences, this demand of value-freedom has been interpreted so that all axiological or normative value terms should be excluded from the language of science. However, for design research the situation is quite different: as we have seen in Section 5, technical norms speak conditionally about values and goals, but the relation between means and ends can be defended in a value-neutral way.

This view agrees with the famous defense of objective social science by Max Weber in 1904.[18] Weber, who accepted the fact – value distinction, held that ultimate or categorical values cannot be proved scientifically, so that they do not belong to the goals or results of scientific inquiry. On the other hand, statements about instrumental value, or relations between given ends and rational means of establishing them, can be defended by empirical scientific investigations.

A related view was defended by Lionel Robbins in his widely read essay on economics.[19] According to Robbins, "economic is the science which studies human behavior as a relationship between ends and scarce means which have alternative uses." But Robbins added that economics is "entirely neutral between ends". This demand of neutrality is misleading, however, as applied sciences typically are interested in socially *relevant* ends.

Design sciences with technical norms of the form (3) can be used for rational planning and decision-making, when the end A is accepted as a basis of social action. The relevant value goal A may be characteristic to the design science: for example, health for medicine and nursing science, profit for business economics, welfare for social policy studies and social work, and peace for peace research. But already the case of medicine shows that for many design sciences the choice and specification of the relevant value goal may be a matter of philosophical, legal, ethical, and political debates. The sources of values of technical norms may thereby be in philosophical arguments, general morality and ethics, empirical value studies, value profiles of institutions and funding bodies of research, and political debates.

This kind of multiplicity of values could be avoided, if moral or axiological realism would hold, so that there are objective goals to be determined by scientific or philosophical arguments. Then the antecedent A could be eliminated from the

18 See Max Weber, *The Methodology of the Social Sciences*. New York: The Free Press 1949. See also Carl G. Hempel, "Science and Human Values", in: *Aspects of Scientific Explanation*. New York: The Free Press 1965, pp. 81-96.

19 See Lionel Robbins, *An Essay on the Nature and Significance of Economic Science*. London: Macmillan 1932.

norm (3) which would be transformed to a simple recommendation (cf. (2)). But such a realist position has its problems, as values are human-made social constructions.[20] A democratic society should be open to free value discourse. In particular, futures studies should allow different value goals for its scenarios, including estimates of the values of future generations.[21]

The technocratic and conservative approach is to accept the value A uncritically, maintaining the status quo. The reformist strategy, exemplified by Karl Popper's "piecemeal social engineering", specifies A with small improvement in social conditions.[22] The emancipatory approach proposes a goal A which is critical of the existing situation and implies radical changes in the social order.[23] In this way, action research and critical social science can be included in the same model of social design science.

The notion of technical norm illuminates also the existence of *policy conflicts* in many fields of study. Disagreement about the best policies X may be due to differences in the knowledge about situation B, in the decision to keep B stable or change it, in the knowledge about the law X&B → A, or in the valuation of goal A. It is important task of philosophical conceptual analysis in applied ethics to distinguish these different sources of disagreement.

7. Examples of applied social sciences

Applied social sciences, their values and organization can be illustrated by examples. The cases show what kinds of sciences have been neglected by philosophers of science.

Richard Titmuss, Professor of Social Administration at the London School of Economics in 1950–73, was pioneer in making social work an academic discipline. He received in 1960 an order from the Governor of Mauritius who wished to know how the population on the island could be controlled. The answer of the Titmuss report was clear: to reduce the need of large families with many children,

20 For a critical assessment of moral realism, see Ilkka Niiniluoto, "Facts and Values – A Useful Distinction", in: Sami Pihlström and Henrik Rydenfelt (Eds.), *Pragmatist Perspectives*. Acta Philosophica Fennica 86. Helsinki: Societas Philosophica Fennica 2009, pp. 109-133.

21 For an account of futures studies as a combination of visionary plans for improving the world and a design science for realizing these goals, see Ilkka Niiniluoto, "Futures Studies: Science or Art?", in: *Futures* 33, 2001, pp. 371-377. Alternative scenarios, which indicate paths from present situations to alternative futures, can be understood as generalizations of the notion of technical norm. For a different approach, where categorical value and ought statements are taken to be empirically justifiable assertions, see Wendell Bell, "Moral Discourse, Objectivity, and the Future", in: *Futura* 28, 1, 2009, pp. 43-58.

22 See Karl Popper, *The Poverty of Historicism*. London: Routledge 1957.

23 See Habermas, *op. cit.*

introduce security by social policy programs.[24] This recommendation can be formulated as a technical norm: if you wish control population growth in poor countries, you should improve social security.

Today *social work* examines the conditions required by people to function and survive day-to-day. The study of individual survival skills and strategies include child welfare, problems facing the youth, pressures in the family, ageing, and marginalized groups like homeless women, HIV-positive, drug addicts, and prisoners. The City of Helsinki and University of Helsinki have together established Heikki Waris Institute as a research and teaching clinic for urban social work.[25] While social work is concerned with a minimal "survival" level of individual human life, the ultimate value premises of *social policy studies* and urban planning is the good of human beings, their quality of life, measured by subjective experiences (satisfaction, happiness) and objective social indicators (basic needs, food, housing, health, wealth, security, and education).

The Nordic model of welfare state is based on the goal of well-being, defined in 1975 by the Finnish sociologist Erik Allardt with three conditions: *having* (material and economic resources), *loving* (human relations), and *being* (self confidence, life politics).[26] Connections to Amartya Sen's account of the quality of life in terms of a fair distribution of capacities or capabilities are obvious.[27] The mean of three value goals is also included the Human Development Index, produced by the United Nations Development Project (UNDP) since 1990: health (life expectancy at birth), education (adult literacy, years of schooling), and living standards (wealth measured by GDP per capita).

The Genuine Progress Index (GPI), proposed by Redefining Progress, adds to GDP other economic factors like income distribution, services outside the market, and costs of negative effects (crime, resource depletion, pollution, loss of wetland). The Happy Planet Index (HPI), published by the New Economic Foundation since 2006, takes seriously the value of environmental protection and sustainable development. It uses the formula: life satisfaction x life expectancy per ecological footprint. These new measures of social progress are today actively discussed by governments in many countries, including the United Kingdom, France, and Finland, but applied research programs with these value goals still wait for their realization.

The City of Helsinki, the Ministry of Education and the University of Helsinki agreed in 1998 about the establishment of six new professors of *urban studies*, and

24 See Richard M. Titmuss and Brian Abel-Smith, *Social Policies and Population Growth in Mauritius*. London: Routledge 1968.

25 Heikki Waris, Professor of Social Policy at the University of Helsinki in 1946–68, introduced social work into the academic curriculum in Finland in the 1950s.

26 See Erik Allardt, "Having, Loving, Being: An Alternative to the Swedish Model of Welfare Research", in: Martha Nussbaum and Amartya Sen (Eds.), *The Quality of Life*, Oxford: Oxford University Press 1993, pp. 88-94.

27 See Nussbaum and Sen, *op. cit*.

later in 2003 the nearby cities of Espoo, Vantaa, and Lahti joined with the Helsinki University of Technology. The fields of the professors cover both descriptive basic research and applied design research: European metropolitan planning, urban history, social policy, urban sociology, urban economics, urban ecology, urban ecosystem, urban technological systems, and urban geography. The underlying values of these studies could be related to the classical ideals of *urbanité* (as opposed to rural life) – elegance, sophistication, politeness, fashion, learning, education, free thinking, public power, close services, interplay of work and leisure, and avoidance of decadence, criminality, poverty, slums, dirt, noise, haste, and loneliness. The City has its own "Helsinki vision", stating that "Helsinki will develop as a world-class innovation and business centre based on the power of science, art, creativity, and good services". The City Planning Department has formulated a "Future City" mission of Helsinki as a multicultural metropolis, a Baltic Sea logistics centre, a European centre of expertise, a world-class business centre. The "official" values of the City are health, safety, and beauty, and additional values include customer-orientation, sustainable development, justice, economy, safety, and entrepreneurship. The statistical office, Helsinki City Urban Facts, promotes strategic decision-making by gathering reliable information.

Brundland's report *Our Common Future* in 1987 made sustainable development as a fashionable theme. In the Johannesburg Summit in 2002 *sustainability* was defined to include environmental protection, economic development, and social development.[28] An interesting example of a new type of research unit, which mixes natural and social sciences, is ICIPE (International Centre of Insect Physiology and Ecology),[29] founded in Nairobi in1970. Its mission is to support sustainable development by the conservation and utilization of Africa's rich insect biodiversity, but at the same time work for human, animal, plant and environmental health. ICIPE aims at improving the overall well-being of communities in tropical Africa by addressing the interlinked problems of poverty, poor health, low agricultural productivity and degradation of the environment.

Department of Philosophy, History, Culture, and Art Studies
University of Helsinki
P.O. Box 24
00014, Helsinki
Finland
ilkka.niiniluoto@helsinki.fi

28 See Taina Kaivola and Liisa Rohweder (Eds.), *Towards Sustainable Development in Higher Education – Reflections*. Helsinki: Ministry of Education 2007.

29 See Liz Ng'ang'a and Christian Borgemeister (Eds.), *Insects and Africa's Health: 40 Years of ICIPE*. Nairobi: International Centre of Insect Physiology and Ecology.

Arto Siitonen

The Status of Library Science: From Classification to Digitalization

Abstract

The essay is concerned with library science as a science of the artificial. This is an area of research that has in recent decades gone through a profound change. The Aristotelian paradigm has made room for an interactice, technologically oriented applied science. This development has transformed the very concepts of book, library and information. Section 2 addresses the concept of book and the history of bookmaking. Section 3 is dedicated to the historical background of libraries. Section 4 concerns the state of the art of modern library science. Section 5 clarifies, how librarians are trained in their profession and which requirements they are expected to satisfy.

1. Introduction

Library science belongs to the class of sciences of the artificial. It is concerned with books, reading and writing. Books are a special kind of artefacts, and thus libraries as collections of these are collections of artefacts. Modern libraries contain alongside books also magazines and newspapers, as well as music and film materials. In addition to that, there are special music libraries and film libraries. In recent decenniums it has become possible to record all information, books and other texts as well as music and films into the form of CD's, i.e. compact discs. This information can also be recorded into computers, and into internet, an international electronic network of data – the World Wide Web, www. The technology behind all these innovations is based on digits, i.e. numerals from 0 to 9. It is called digital technology, which is a branch of electronic technology.

This development implies that books and libraries are undergoing a transformation unseen before. There are real books and libraries; on the other hand, there are virtual books and virtual libraries. This transformation concerns the profession of librarians as well. The competence that is presently required from librarians, exceeds by far that expected from previous generations of them. Accordingly, the training that librarians need for learning the skills that are necessary in order to succeed in the profession, is much more sophisticated and demanding than it used to be in history. Also the very concepts of library and book have changed in this process.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_22,

The method of a traditional librarian can be characterized as the *Aristotelian* one. Books are classified according to some adopted categories, and within these the authors' names determine their ordering. There are sub- and super-classes. This taxonomy helps the customers of a library to find the books that they would like to read or inspect. In modern libraries, the procedure is still observed. However, it is no longer sufficient. Many works of a given library, perhaps even all of them, can be found also in virtual space, as digitalized into electric form; in this case, we speak of *e-books* (and *e-magazines*) that are put into internet. In fact, internet has revolutionized what we understand by books and literature. This is due to the fact that one may not only read a book, but also interact with it in "the Net". What is a book, if it is no longer a closed work, a finished artefact? The identity of works becomes an open question. This raises corresponding copyright problems – which, moreover, appear to put the concept of copyright, as customarily understood, in question. Really, a *paradigm change* has taken place: we have proceeded from classifying stable books into digitalizing them – from physical things to codes. The Aristotelian classificatory approach has been complemented, or even replaced, by digitalization.

The so-called *digital revolution* can be characterized as follows: all information can be packed into digital form – not only books and other texts but also video and audio materials (images and sounds; i.e. pictures, films, speech, music). One may cooperate (play) with this material. Digital revolution has fed techno-optimism; but it may also give reasons for pessimistic scenarios. What happens to libraries? What happens to books? How can we cope with the virtual world which clearly differs from the real world as well as from the imaginary worlds? We have been at home with the latter ones, but do we really master the virtual world? Can we do that? Moreover, how vulnerable are our digitalized treasures?

The following topics will be discussed below: books, libraries, library science and education of library professionals. These are put into historical context. A reflection follows as to the cultural significance of literacy. Books have always served the tasks of entertaining and educating readers. Books and libraries have also had the twin roles of research objects and research sources. Due to this, they have served science.

2. On books

Books are, first of all, concrete physical things made of paper (or originally bast or reed; cf. the Greek word 'papyrus') and skin or some other cover material. They contain alphabetic symbols and other signs, occasionally also pictures, photographs, drawings, diagrams – something to be read or inspected. Books give information that is meant to teach and entertain readers. The titles of the books give potential readers hints as to the topics that are concerned in them. That helps

readers to identify and classify them. Books are artefacts that have been produced in order to preserve information, whether textual, pictorial or auditive. Reading a book and writing a book are complementary operations, contrary skills. It is needless to say which of these two operations is the more demanding one.

In his work *Writing. The Story of Alphabets and Scripts*,[1] Georges Jean traces the origin of books. First of all, reading requires what he calls "The Alphabet Revolution".[2] Due to the invention of alphabet it was "possible to write anything at all by using only about thirty signs".[3] Before this, reading and writing had required a mastery of "a large number of signs or characters".[4] The roots of this revolution were the Phoenician writing system and the Greek script. In principle, all writing systems are based on the idea that speech has to be co-ordinated to signs that represent and preserve the spoken words. Words in written forms consist of signs. One may describe this physically and physiologically as follows: written words correspond to louds that come from the mouths and throats of speakers and vibrate in the air, to be captured by the ears of listeners.

The next step in the evolution of books is characterized by Georges Jean as "From Copyists to Printers".[5] In order to make various works accessible to readers as well as to libraries, these had to be copied by writing the text anew. Copyists were occupied with this task in ancient times and middle ages. Copying was an invaluable albeit time-consuming activity. A revolutionary, decisive improvement was the invention of printing. This was due to Johann Gutenberg, who in 1448 printed in Mainz a book on astrology. The printing technology made it possible to dramatically replicate the copies of a book. Methods of printing improved further; see, for example, litography since 1796, colour printing since 1880, offset-method and computer-aided printing in the 20th and 21st centuries.

Georges Jean adds further steps to the process of book manufacture: "Bookmakers", whose task is to give the finishing touch to the works under preparation, e.g. through binding,[6] and "Decipherers", who may be needed in order to render various ciphers readable and understandable.[7] The task of the latter is, paradoxically, to "decipher the indecipherable". According to Georges Jean, "the first, and surely the most inspired of all", was Jean-Francois Champollion (1790–1832), who managed to open the hieroglyphs of the famous Rosetta Stone. One may also mention Michael Ventris, who deciphered the Cretan script, the so-called Linear B, during 1950–52.

---

1 Jean Georges, *Writing. The Story of Alphabets and Scripts.* Translated from the French by Jenny Oates, New York: H. N. Abrams 1992.
2 *Ibid*., pp. 51-71.
3 *Ibid*., p. 52.
4 *Ibid*.
5 *Ibid*., pp. 73-96.
6 *Ibid*., pp. 97-116.
7 *Ibid*., pp. 117-126.

Georges Jean closes the Decipherers chapter with the subtitle "There Are Still Many Undeciphered Signs".[8] Among these are Linear A and the Phaistos disk; neiher "has as yet given up its secrets".[9] "The riddle of the unsolved scripts, the brilliance of those who invented them, and the genius of those working to crack the code all continue to fascinate us".[10]

In the same way as philosophers are those who love wisdom, bibliophiles are lovers of books and collectors of these. The Greek word *biblion* refers to a paper, a scroll, a letter. Closely related to this word is the title "Bible", distinguishing an authoritative book, especially Christian scriptures in the Old and New Testament, from other, so-called mundane books. The history of translations of the Bible has been fruitful for the development of the languages into which it has been translated. Correspondingly, translations of ancient texts (Homeros, Hesiodos, Plato etc.) have furthered the world culture.

The concept of book has its typical, albeit somewhat fuzzy boundaries. There is a saying that something is "of a book's length", i.e., wide and large enough in order to be called a book. For instance, a distinction is drawn between novels and short stories (or *nouvelles*). Various comparisons can specify when a story forms a genuine novel, and when not. A corresponding definition task is given, when a scientific book is distinguished from a scientific article.

On the other hand, a definitely clear distinction is to be drawn between concrete books and so-called *web books*. The former ones consist of atoms and are physical things, while the latter ones are virtual and require, in order to be read, electricity. From the *mental* point of view, there are differences between the attitudes and experiences of those who are reading a "genuine book" and those who read an "e-book" – whether with the help of a reading pad, or from the internet. *Noetically* considered, there is no difference whatsoever: the same abstract content is involved in these diverging processes of reading.

However, it is possible to manipulate the contents. In this respect, internet has in fact revolutionized the very concepts of book and literature. It renders a new format for old contents. We have paper books, and we have online versions of these. Due to the latter, it is possible to consider books as open codes. This makes books interactive. We can play games with books, transform them after they have been published in the internet. This has been shown, for instance, by the young Russian writer Dmitri Gluhovski, who in 2002 published his novel *Metro 2033* as an audiobook in the World Wide Web. His story belongs to the genre of *dystopia* literature; its theme is survival of mankind in a global cataclysm. There is also a paper version of this work available in the "real world;" moreover, a video-game of the book has been developed in the internet. The latter extends the novel – and

8 *Ibid*., p. 126.
9 *Ibid*.
10 *Ibid*.

also its readership: one may take part to that game and accordingly transform the original story.[11]

Presently, there are over 200 millions persons in the world who are using the World Wide Web. One may wonder which effects this "virtual revolution" will have upon the cultures of reading, writing and publishing, as well as upon society and human beings in the long run.

A concrete guide to the issues of reading, writing and publishing books, to the skill of practising philosophy, and to the problem of finding books in the internet has been written by the Swedish philosopher Sven Ove Hansen: *Verktygslära för filosofer* ('Organon for Philosophers'), published in the year 2010 (originally in 2009) in Stockholm.[12] It is a combination of textbook and reference work. Hansen's work has not been translated into English, but it would certainly deserve that.

3. On libraries

Traditionally, libraries are private or public collections of books or manuscripts. They are likewise places or buildings in which such collections are kept. Private persons may be collectors of books who occasionally buy or receive these and put them in any order or disorder at their home, whereas public libraries systematically organize their collections according to fixed principles. These determine the classificatory schemata that are adopted in order to place books on the shelves. This is done in alphabetical order that runs from A to Z, for each subclass.

Thus, for instance, there may be a class for books on sport, its subclasses being called according to various disciplines of sport. "Athletics" can be the name of one of the subclasses. In a sport school library, there may even be a finer distinction of this subclass, in which track and field disciplines form their respective sub-sub classes. Athletics is concerned with running, jumping and throwing, all of these having several branches as, for instance, 100 meter running, 110 meter hurdles, 3000 meter steeplechase, etc.

Classificatory principles of books vary between various libraries. In a village library, there used to be a class named: "large books". In this case, it is not the content nor the title of the respective works that determine their placement in the library, but rather their physical size.

The word "library" stems from the Latin word *librarium* that indicates a place, a dwelling of books, a home for books (*libri*); meanwhile in Latin *liber* is book. Public libraries are communal institutions. They contain collections of books and other items that their customers can loan for reading, watching or listening pur-

11 Cf. also *Metro 2034* – Gluhovski's own continuation of the story, *www.like.fi.kyet/metro* (access on 11.2.2011).

12 Sven Ove Hansen, *Verktygslära för filosofer* ('Organon for Philosophers'). Stockholm: Thales 2010.

poses during a certain amount of time. University libraries are open for students and staff. Various organizations like schools, hospitals, firms and clubs often have book collections that stay at the disposal of their members. Some libraries contain so-called special collections, from which customers can loan documents or at least photocopies of documents, or use digital services in order to obtain these. There are also children's libraries, as well as library rooms that are dedicated to children's books.

Perhaps the most famous and legendary library in history has been that of Alexandria – a city which was founded in Egypt by the emperor Alexander the Great in the year 331 BC. The library itself was founded in ca. 295 BC by Demetrius Phalereus, a pupil of Aristotle's. This library was the largest and the most remarkable one in ancient times. It was in fact a teaching, learning and research center, that was built according to the model of Aristotle's school *Lykeion*. Alexandrian library collections consisted of approximately half a million papyrus rolls, called *bibliothekai*. Among others, the library contained Demosthenes' speeches. It did not only collect literature but also produced it. The foundations of physics, astronomy, geology, biology, medicine, mathematics, study of grammar, textual criticism and linguistics were established in this library. Kallimakhos, a Greek grammarian and poet, organized there the compilation of national bibliography.

Alexandrian library was devastated many times. *Mouseion*, the place that was devoted to research, burned down in 47 BC. Soon after that the library of Pergamon was transferred to Alexandria. It was destroyed in the civil war under the Roman emperor Aurelianus in the late 3rd century AD. Arabs conquered Alexandria in 642. Today, it is an Egyptian city.

In the present Alexandria, the famous library is experiencing a reincarnation. It was founded in the year 2002. The goal of this library, called *Bibliotheca Alexandrina*, is to promote the ideal of "Universal Access to Human Knowledge". This contains the project of digitalizing all library collections in the world, and creating an Open Content Alliance (OCA). Connected to this plan is a film archive that is under construction. The new library of Alexandria is committed to the spirit of openness, renovation of society, as well as international dialogue, tolerance and cooperation. The way from the ancient Alexandrian library to the present rebuilt Alexandrian library can be characterized as a development from clay tables to papyrus and feather, and from these to internet.[13]

Among famous libraries of the world, one may also mention two remarkable national libraries: the British Museum Library in London, founded in 1753,[14] and the Library of Congress in Washington DC, USA, founded in 1800. The Congress Library contains, among others, an invaluable collection of jazz music record-

13 Cf. *www.britannica.com/.../Library-of-Alexandria* (access on 11.2.2011). Cf. also Hannu Pesonen, "Bittiajan ihme" ('The Miracle of the Bit Period;' cf. Binary Digits), an article in the Finnish Magazine *Suomen Kuvalehti* 29, 2011, pp. 40-45.).

14 Cf. *www.britishmuseum.org* (access on 11.2.2011).

ings.[15] During recent years, this collection has been digitalized. The Library of Congress had until recently the world largest collection of printed works, but presently it has been overtaken by the reborn Alexandrian library.

Various catastrophes like wars and earthquakes have devastated human cultural monuments, including libraries and books. Let us add a further example to the devastations that took place in Alexandria. The Austrian author Stefan Zweig in his work *Sternstunden der Menschheit* describes twelve historical occurrences that "changed the world" – as the saying goes. He devoted one of these episodes to the conquest of Byzantium that took place in May 29, 1453. This was due to Sultan Muhammad, who led the Turkish army. The catastrophe included the annihilation of innumerable and immeasurably valuable books in the libraries of that city. As Zweig says, books that contained the wisdom of centuries, the immortal richness of Greek thinking and poetry that was meant to be preserved for ever, were burnt or thrown away.[16]

Alongside historically remarkable libraries, there are also fictitious libraries. One of the most curious ones of these is described by Jorge Luis Borges in his work *Ficciones* from the year 1956, in the story *The Library of Babel*. It is to be noted that Borges had been director of the National Library of Argentina – an ideal profession for a writer. In the story he speaks of "a general theory of the Library"[17] and of "the fundamental law of the Library".[18] Borges even axiomatizes his theory. There are two axioms: "The first: The Library exists *ab aeterno*. No reasonable mind can doubt this truth, whose immediate corollary is the future eternity of the world".[19] "The second: *The number of orthographic symbols is twenty-five*".[20] Borges also addresses the problem of infinity, ending up with the claim that the Library is "*limitless and periodic*".[21] Indeed, this library clearly exceeds all existing libraries – and, one may claim, the imaginative ones as well.

15 Cf. Marshall Stearns: *The Story of Jazz*. Oxford: Oxford University Press 1956.

16 Cf. The chapter "Die Eroberung von Byzanz", translation from p. 64 of that book: Stefan Zweig, *Sternstunden der Menschheit. Zwölf historische Miniaturen.* Frankfurt/Main: Fischer Verlag 1985 (1st ed. 1929). The original text is as follows: "… die Bücher, in denen die Weisheit von Jahrhunderten, der unsterbliche Reichtum des griechischen Denkens und Dichtens bewahrt sein sollte für alle Ewigkeit, verbrannt oder achtlos weggeworfen".

17 Jorge Luis Borges, *Ficciones*. Ed. Anthony Kerrigan, New York: Grove Weidenfeld 1962, p. 81.

18 *Ibid*., p. 82.

19 *Ibid*., p. 80.

20 *Ibid*., p. 81.

21 *Ibid*., p. 87.

4. On library science and informatics

Library science is the discipline that concerns the branch of librarianship. Traditionally, this science has two basic aims, theoretical and practical. The previous one is devoted to the study of history of libraries, while the latter concerns technical organization and maintenance of libraries.

Modern library science belongs to *applied sciences*. It is closely connected to the *theory of information*, as is indicated by its former name "library science and informatics". Presently, this area of science is called "information research". Its aim is to inquire into the production, transmission, organization, acquisition and use of information. In logic and mathematics, information is defined quantitatively as a certain measure; the basic unit of information is called a "bit". Unlike matter, bits do not consist of atoms. They are weightless, and they travel with the speed of light. A bit can be considered to be a 1 or a 0, and this binary representation may be interpreted in innumerable ways. Thus, in principle, all media can be transformed into the digital form.[22]

Theory of information is a quantitative study that is concerned with transmission of information by signals. In order to speak of knowledge, a given piece of information has to be true and justified. Accordingly, the concepts of information and knowledge have to be kept apart. This is required by the theory of knowledge: whereas information may be true or false, truth is a necessary condition of knowledge.

From the point of view of pragmatics, messages always have a sender and a receiver. Decisive questions here concern the *intelligibility* of messages and the ability of *understanding* messages. Senders fail in their task if their messages are incomprehensible, whereas lack of understanding, as well as misunderstanding, are cases in which receivers fail. The issue of understanding is closely related to that of interpretation. *Hermeneutics* is the area in which the operations of understanding and interpretation are studied. Its results can be profitably applied to digital technology, whereas the technology that has been developed in the Media laboratories helps us in our enjoyment of audio, video and textual information. According to Negroponte, these Media Labs have brought "a deep understanding of both the human sciences and the sciences of the artificial".[23] He also claims that the lines between art and sciences are fading: "… we are finally moving away from a hard-line mode of teaching…toward one that is more porous and draws no clear lines between art and science".[24]

What is called "library science" is actually a happy combination of art and science. The *art* lies in literature, in the culture of writing and reading and of preserving our literary tradition and its values. All of us, as human beings, have grown

22 Cf. Nicholas Negroponte, *Being Digital*. New York: Alfred A. Knopf 1995, p. 12 ff.
23 *Ibid*., p. 234.
24 *Ibid*., p. 220.

from illiteracy to literacy, because the necessary cultural skills have been taught to us. *Science*, in turn, lies in a strict, formal theory of messages, their transmission and their senders and receivers. This theory forms the meta-level to the study of reading and writing. On the other hand, the object level of library science concerns books and libraries themselves.

It is interesting to compare the present state of library science to its *pre-digital states*, especially to the last phases before the very turning point that propelled digitalization. A good example of this is a seminar in Helsinki that was held on November the 24th and 25th, 1984. Its theme was 'Libraries and Science'; the proceedings were published in 1985. The first article in the publication was by Ilkka Niiniluoto on the topics 'Library, Science and Library Science'.[25] According to him, the branch 'library science and informatics' has recently established itself and is searching for its identity. The main themes of his article are: interaction between libraries and science, library as a research object, levels of information research, descriptive library science, library science as science of planning, and library politics.[26]

Niiniluoto emphasizes that a large part of libraries' books are scientific works that can be used for learning purposes; books are products of culture as well as objects of research; libraries can be seen to "constitute" science. Scientists have to learn to use library services in their teaching and research work; correspondingly, every library professional should master the basic knowledge of philosophy of science. According to a known saying, libraries are "the memory of mankind". Ever since the Babylonian, Egyptian and Greek cultures flourished, libraries have carried the written tradition forward and thus made cultural evolution possible. The tasks of modern library have expanded, because it has to take care of information services. Information is considered to be either physical, syntactic, semantic, or pragmatic. Library science belongs to applied sciences. It was professionalized when the world's first chair of library science was founded at the University of Göttingen in 1886. Already long before that, there was a textbook on this science available: Martin Schrettinger's *Versuch eines vollständigen Lehrbuchs der Bibliothekwissenschaft* (published in 1808).

Niiniluoto draws a distinction between library science and *library philosophy*. The latter is concerned with the general value premises and main goals of library institution.

Since the publication of Niiniluoto's article, 27 years have elapsed. Meanwhile, Negroponte's work *being digital* appeared in 1995. This work reported the state of the art that had been acquired until its publication. Thus, there lies a ten years progress (from 1985 to 1995) during which digital technology revolution-

25 Ilkka Niiniluoto, "Kirjasto, tiede ja kirjastotiede", in: Tuula Haavisto (Ed.), *Kirjastot ja tiede*. Helsinki: Kirjastopoliittinen yhdistys 1985, pp. 5-22.

26 *Ibid*. For the expression "science of planning", see Herbert Simon's work on the *Models of Discovery*. Dordrecht: Reidel 1977. On planning processes, see Simon´s *Models of Discovery*, pp. 226 f.

ized books and libraries as well as our ways of thinking of these and handling them. Presently, we have enjoyed the technology of digitalization 17 more years (from 1995 to 2012) and begin to see its effects even more clearly than these were reported by Negroponte.

5. On librarians' education and challenges

The profession of librarianship has radically changed during the recent years. The requirement of constant training concerns all librarians – but also their customers. Lending books from a library often involves the use of computers and data bases. For library professionals, this is elementary. Presently, becoming a librarian presupposes a duly qualified examination in library science and information theory, as well as a further examination that concerns information and library services. A librarian has to master the history of libraries, the organization of data, and be able to evaluate the architecture of information.

Moreover, a librarian's studies include scientific communication and bibliometrics, planning and evaluation of web services and games, language technology and methods of data search, library services, their evaluation and improvement, as well as selling and production of data resources. The studies contain also courses on interactive media, society and culture, cultural-sociological viewpoint to the library institution, approach to web publishing, research concerning the use and the users of interactive media, expertise of library-and-information branches and leadership, experimental data search, theories of game research, the history and the present state of information and media systems.

As one may see, the program comprises technological, natural scientific, social scientific and humanistic studies. It has a thoroughly interdisciplinary character. The profession of librarianship has really been radically transformed. Traditionally, a librarian's main task, alongside serving the customers, was to classify and catalogize the works that comprised the library in which she or he was working. Classification and service are still mandatory, but the required expertise contains layers of tasks that were previously unknown. In short, library science has undergone a thorough *paradigm change*.

It should not be an exaggeration to characterize library and information specialists hard-core professionals; they have had a tough education and are expected to be constantly open-minded for the new challenges that they are bound to meet in their profession. They serve and preserve the culture of literacy. For the concept of 'culture of literacy,' see the work by Wlad Godzich.[27] Today's literary culture does not exclude but includes technological skills.[28]

27 Cf. Vlad Godzich, *The Culture of Literacy*. Cambridge (Mass.): Harvard University Press 1994.

28 The source of the above report is the decree for librarians, *Kirjastoammatillisen henkilöstön kelpoisuusvaatimukset 1.1.2010-* ('Proficiency requirements for library profes-

6. Conclusion

The concept of book has gone through modifications; and, together with it, that of a library as a collection of physical artefacts. Nevertheless, we can still speak of books, libraries and literature – in a *concrete* sense as artefacts and buildings, and also in an *abstract* meaning as mental and conceptual contents, or in a *virtual* meaning as digitalized bits of information.

We have in fact managed to enhance our possibilities to be participants in literacy culture. We need more than ever before literary education; fortunately, that is available in plenitude. We are no longer only recipients of literature but also active actors in its production and acquisition. Moreover, our librarians are better educated and equipped than ever before.

It is possible to contrast the sciences of the artificial, such as library science, to cultural and social sciences. However, books and libraries are constituting factors in culture; and books socialize their readers. Thus, cultural and social sciences rely on literature and its availability. Books, libraries and library science enhance our possibilities of understanding foreign cultures and societies – as well as ourselves. Some scruples remain concerning our reliance on technology. There reigns a widespread *techno optimism* that is especially connected to computers, internet and expanding digitalization. New software is constantly introduced, and hardware (machines) gets replaced by new models. However, various gadgets and programs appear to determine and even dictate the way how we think and how we organize our practises of writing and reading. This development may backfire. It is true that real, genuine, physical books are still manufactured and sold in bookshops as well as delivered to libraries. A sceptic may wonder, how long these practices will continue. It is possible to publish books only in the internet and guide customers to recall these from there and eventually to print them on their own. Such a "do it yourself"-mentality is indeed widely distributed. This is unfortunate. Nothing can compensate the original enjoyment of books and reading. Buying or lending a genuine book and holding it in hands, feeling its paper and smelling its scent, turning over the leaves, and reading it are irreplaceable experiences.

A further example is the replacement of film rolls by bits. This practice may in some cases lead to unexpected difficulties. Digital copies need a secret code that functions only in limited places with special devices and within certain interval of time. The maintenance of films in archives is a risky business. If film theaters will be wholly digitalized, film projectors disappear. Digitalization projects require a translation of film copies into a new format, and this may take years to materialize. A real threat is that old collections cannot be presented at all. Furthermore, the wording on cinema (subtitles that are added into foreign films) is endangered. There is no guarantee that this would be preserved when a film is shown anew. Es-

sionals', put forward by the cabinet of Finland since the beginning of the year 2010'; additional paragraph in 2011). Helsinki, February 3, 2011.

pecially, small languages are hard hit. Thus, audio-visual archives are confronted with serious challenges.

Indeed, digitalization belongs to the tough issues. It is not always a solution to problems but can also raise problems that require solution. A certain amount of scepticism is at place in order to mitigate exaggerated enthusiasm for technology. However, its representatives might ask the sceptics: do you really want to return to typewriters or even to feather, and to handwritten books? A suitable answer would be: of course not, but let us be sober with technology, enjoying real books as well as using digital services. As to feather, it belongs to history, but what is still recommendable is to use fountain pen while writing the first version of a manuscript. This is much more enjoyable than working under the premises of a computer. Moreover, also calligraphy can be recommended.

Department of Philosophy
University of Helsinki
Unioninkatu 40
00014, Helsinki
Finland
arto.siitonen@helsinki.fi

Paolo Garbolino

The Scientification of Forensic Practice

Abstract

Forensic science is traditionally defined as the science of individualization, but claim that forensic scientists are able to achieve conclusions of individualization has been criticized in recent years. Many scholars hold that perfect identification of a person or object as the source of trace or mark is unachievable and that opinions about the source are always probabilistic. From the very beginning, forensic science met statistics and probability theory, and its history provides a good case study of what Bernard Cohen has called the "probabilizing revolution" in science. The emergence of DNA typing has set up a major challenge for forensic practice, opening the door to the use of advanced statistical methods and putting in question the scientific status of traditional forensic methodologies.

1. The science of identification

Forensic science is defined as the *science of individualization* by practitioners and most of theorists: "Criminalistics is the science of individualization" claimed a seminal paper in 1963, and an up-to-date textbook maintains that "The state of practice of forensic science is that examiners do provide opinions of individualization".[1] Individualization is the process of addressing the issue whether or not the source of a particular trace is a single object or person.

For example, a window has been broken during the commission of a crime and a blood stain is found on a fragment of glass of the broken window. A suspect is found with fragments of glass on his clothing. Several fragments are taken from the crime scene and from the suspect's clothing and a comparison is made between the two samples of fragments on the basis of their refractive index measurements. A sample of *DNA* material is taken from the blood stain found at the crime scene and another sample of *DNA* is taken from the suspect and a comparison is made between the two on the basis of their *DNA* profiles. The task of forensic scientists is to answer questions like: "do the fragments of glass found on the clothing of the

1 Paul Kirk, "The Ontogeny of Criminalistics", in: *Journal of Criminal Law, Criminology and Police Science* 54, 1963, pp. 235-238; Keith Inman and Norah Rudin, *Principles and Practice of Criminalistics: The Profession of Forensic Science*. Boca Raton: CRC Press 2001, p. 148.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_23,

suspect come from the broken window?", and "does the bloodstain found at the scene of the crime come from the suspect?".

Forensic scientists distinguish *individualization* from *identification*, where identification is understood to mean the narrowing of possible sources of a trace not to a single person or object but to a class of objects, as when an eye-witness testifies that she saw a black Toyota in a particular place at a particular time.[2] The term identification will be used throughout this paper for the following reason: the traditional claim that forensic scientists are able to achieve conclusions of individualization has been strongly criticized in recent years, and many scholars hold that perfect identification of a person or object as the source of trace or mark is unachievable and that opinions about the source are always probabilistic.[3] Some of them maintain that identification should be conceived as a *decision* rather than a *conclusion*, and argue for the need to apply the methods of statistical decision theory. This issue is part of the epistemological problem whether, and to what extent, it is possible to transform the arts of comparing tool marks, fingerprints, fibers, handwriting and organic liquids into a scientific practice.

2. The probabilization of identification

Hunters, sailors, art *connaisseurs* and physicians have been looking for centuries to natural signs, and their observations were conceived of as reading testimony: things can bear witness and their testimony is the testimony of nature itself.[4] During the 19th century, the science of identification made its first steps in the field of legal medicine: Joseph Bell (1837–1911), a celebrated forensic physician from

2 David Kaye, "Identification, Individualization and Uniqueness: What's the Difference?", in: *Law, Probability and Risk* 8, 2009, pp. 85-94.

3 David Stoney, "What Made us Ever Think we Could Individualize Using Statistics?", in: *Journal of Forensic Science Society* 31, 1991, pp. 197-199; Ian Evett, "Establishing the Evidential Value of a Small Quantity of Material Found at the Crime Scene", in: *Journal of Forensic Science Society* 33, 1993, pp. 83-86; Bernard Robertson and Anthony Vignaux, *Interpreting Evidence. Evaluating Forensic Science in the Courtroom*. Chichester: Wiley 1995; Michael Redmayne, *Expert Evidence and Criminal Justice*. Oxford: Oxford University Press 2001; Franco Taroni, Colin Aitken and Paolo Garbolino, "De Finetti's Subjectivism, the Assessment of Probabilities and the Evaluation of Evidence: A Commentary for Forensic Scientists", in: *Science and Justice* 41, 2001, pp. 145-150; Philip Dawid, "Bayes's Theorem and Weighing Evidence by Juries", in: Richard Swinburne (Ed.) *Proceedings of the British Academy: Bayes's Theorem*. Oxford: Oxford University Press 2002, pp. 71-90; Aitken and Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*. 2nd ed., Chichester: Wiley 2004.

4 See Ian Hacking, *The Emergence of Probability*. Cambridge: Cambridge University Press 1975; Carlo Ginzburg, "Clues: Roots of an Evidential Paradigm", in: Carlo Ginzburg, *Clues, Myths, and the Historical Method*. Baltimore: The Johns Hopkins University Press 1989, pp. 96-125.

Edinburgh, was the real world model of Sherlock Holmes character. Edmond Locard (1877–1966), a French physician who started the first police laboratory in 1910 in Lyon, stated what is considered by practitioners and theorists the basic heuristic principle of forensic science, known in the literature as *Locard's exchange principle*, saying that every physical contact leaves a trace: "either the perpetrator leaves marks of his intervention on the scene, or by an inverse action, he takes traces of his actions on his body or clothes".[5]

Signs of nature began to be read not only as the most faithful witnesses of people's actions, but also as the silent and sure witnesses of personal identity. The measurement of physical human traits, anthropometry, was conceived of as the solution of what had become a very important social problem: personal identification.[6] Forensic photographs were not sufficiently reliable due to the lack of a standard for taking them and of an adequate system of classification. The problem of providing such a system was first addressed by Alphonse Bertillon (1853–1914) and eventually solved by Francis Galton (1822–1911). Bertillon proposed to use measurements of somatic characteristics as height and bones to prove personal identity: if a match with data could be found, taking into account a table of "acceptable" measurement errors, the identification was then completed by examining photographs and other physical marks such as tattoos and scars. The system, called *Bertillonage*, had some important drawbacks: it was a purely comparative system that could not help in deciding whether a particular person had been in a particular place; the measurement errors were based upon what we would call today a *convenience sample*, and, above all, the frequencies of somatic traits in the population were unknown, so that the probability of another person having the same set of measurements, what is called today the *random match probability*, was unknown.[7]

Fingerprints were the solution to the problem of personal identification proposed by Francis Galton.[8] In his laboratory in South Kensington he had been collecting for many years a large bulk of anthropometric measurements for his main research field, biological heredity, and when he applied those data to a scrutiny of *Bertillonage*, he was immediately led to two major scientific breakthroughs: the formulation of a classification system for fingerprints and the concept of *statistical correlation*.[9] The very empirical data that made Galton aware of the sym-

5 Edmond Locard, *L'enquête criminelle et les méthodes scientifiques*. Paris: Flammarion 1920, p. 139.

6 Ginzburg, *op. cit*., pp. 119-120.

7 The *random match probability*, which is the probability that a person in the relevant population, different from the suspect, shares a particular feature, e. g. the *DNA* profile, is based upon knowledge that there is at least one person, the suspect, that has that feature, and therefore can be different from the relative frequency of that feature.

8 Francis Galton, *Finger Prints*. London: Macmillan 1892.

9 For the story of Galton's discovery of statistical correlation, see Theodore Porter, *The Rise of Statistical Thinking 1820-1900*. Princeton: Princeton University Press 1986;

metry between regression and co-relation provided him the reason to argue for the superiority of fingerprint analysis over Bertillon's method: the measures of the limbs, which Bertillon treated as if they were independent, are "undoubtedly correlated".[10] On the contrary, fingerprint features are independent from all other bodily measurements, so that they allow much safer identification than any other anthropometric method.

In fingerprint analysis the main cause of error is that the original mark from the scene, what is called the *latent print*, may be partial and of poor quality. Moreover, the main instrument of observation is the human eye, and fingerprint identification is still today a subjective judgment given by an expert. Galton did not distinguish between the probability of error properly speaking, that is, the probability of a false positive, and the random match probability, the probability that two different and unrelated individuals in the relevant population have the same fingerprints. The failure to make the distinction is due to the hypothesis, held with certainty by Galton and by almost all fingerprint experts even today, that fingerprints are unique, but this is an empirical hypothesis, supported at best by an inductive argument of the same kind of the "all swans are white" argument.[11]

The most influential probabilistic model for latent fingerprints was put forward in 1911 by Victor Balthazard (1872–1950), professor of forensic medicine at the Sorbonne. Balthazard, like Galton, did not distinguish between measurement errors and random match probabilities, and his model was built upon highly questionable hypotheses: the papillary surface was divided into 100 equal squares (points), each one of them assumed to contain one fingerprint trait out of four basic types, each type was assumed to have the same *a priori* probability of being observed in one square, and traits were assumed to be independent one from the other. Given these hypotheses, the probability of finding a match by chance in n points is $(1/4^n)$, and he concluded that, with $n = 17$, the probability was small enough to yield a practical certainty that no other person exists in the entire world population with 17 matching points.[12]

Stephen Stigler, *The History of Statistics. The Measurement of Uncertainty before 1900.* Cambridge (Mass.): The Belknap Press of Harvard University Press 1986. Galton's system was developed in a full workable method by Sir Edward Henry (1850-1931), and the so called Galton-Henry Classification System has been used in most English-speaking countries up to the 1990s, when automated fingerprint identification systems were introduced.

10 Galton, *Ibid*. p. 157. On Galton and fingerprint see: Stigler, *Statistics on the Table. The History of Statistical Concepts and Methods*. Cambridge (Mass.): Harvard University Press 2002, pp. 131-140.

11 See on this point Kaye, "Identification, Individualization and Uniqueness: What's the Difference?", *Ibid*.; and Simon Cole, "Forensics without Uniqueness, Conclusions without Individualization: The New Epistemology of Forensic Identification", in: *Law, Probability and Risk* 8, 2009, pp. 233-255.

12 Victor Balthazard, "De l'identification par les empreintes digitales", in: *Comptes Rendus des Séances de l'Académie des Sciences* 152, 1911, pp. 1862-1864. On Baltha-

Balthazard's model is at the origin of the 17-points rule adopted by scientific police in France and Italy for fingerprint identification, and the 16-points rule that has been used in England up to 2001: 16 or 17 coincidence points are deemed to be sufficient for allowing experts to give in court positive statements of identification. United States and Australia adhere to a less conservative 12-points rule. Balthazar's assumptions have been extensively criticized and other models have been proposed, but there is still no consensus about them.[13]

The first historical example of the use of probability to assign a probative value to evidence in court was the witness given by the Harvard mathematician Benjamin Peirce and by his son Charles Sanders in the Howland case in the 1860s: the problem was the authenticity of the signature of Silvia Howland's testament. This historical case is a good example of the subtleties of probabilistic reasoning applied to forensic evidence: either the two Peirce made an implicit, and unwarranted, assumption of *equal prior probabilities* for the prosecution hypothesis and the defence hypothesis, or their reasoning is an instance of the so called *prosecutor's fallacy*, which consists in taking the likelihood of the defence hypothesis (in this case very small) as its *posterior probability*.[14]

From the very beginning, forensic science met statistics and probability theory, and its history provides a good case study of what Bernard Cohen has called the "*probabilizing revolution*" in science, that is, "the introduction of probability and statistics into areas that have undergone revolutionary changes as a result".[15] Forensic science, in its way to "scientification", needed to collect statistical data, figure out probabilistic models, and find a good way for estimating and commu-

---

zard's model see: Franco Taroni, Christophe Champod and Pierre Margot, "Forerunners of Bayesianism in Early Forensic Science", in: *Jurimetrics Journal* 38, 1998, pp. 183-200.

13 Taroni, Champod and Margot, "Forerunners of Bayesianism in Early Forensic Science", *op. cit*., pp. 197-198; David Stoney and John Thornton, "A Critical Analysis of Quantitative Fingerprint Individuality Models", in: *Journal of Forensic Sciences* 31, 1986, pp. 1187-1216; Stoney, "Measurement of Fingerprints Individuality", in: Henry Lee and Robert Gaensslen (Eds.), *Advances in Fingerprint Technology*. Boca Raton: CRC Press 2001, pp. 327-387; Christophe Champod, Chris Lennard, Pierre Margot and Milutin Stoilovic, *Fingerprints and Other Ridge Skin Impressions*. Boca Raton: CRC Press 2004.

14 See Paul Meier and Sandy Zabell, "Benjamin Peirce and the Howland Will", in: *Journal of the American Statistical Association* 75, 1980, pp. 497-506. William Thompson and Edward Schumann, "Interpretation of Statistical Evidence in Criminal Trials. The Prosecutor's Fallacy and the Defence Attorney's Fallacy", in: *Law and Human Behaviour* 11, 1987, pp. 167-187. The fallacy is a particular instance of the *fallacy of the transposed conditional*: Persi Diaconis and David Freedman, "The Persistence of Cognitive Illusions", in: *Behavioral and Brain Sciences* 4, 1981, pp. 333-334.

15 Bernard Cohen, "Scientific Revolutions, Revolutions in Science, and a Probabilistic Revolution 1800-1930", in: Lorenz Krüger, Lorraine Daston and Michael Heidelberger (Eds.), *The Probabilistic Revolution. Volume 1: Ideas in History*. Cambridge (Mass.): The MIT Press, pp. 23-44, especially, 40.

nicating to laymen, judges and jurors, the weight of forensic evidence, which is unavoidably probabilistic in nature.

3. Poincaré's Rule

Alfred Dreyfus was an officer in the French Army who was accused in 1894 of selling military secrets to Germany. The prosecution presented a document, called the *bordereau*, admittedly written by Dreyfus himself, pretending it contained ciphered messages. One of the argument put forward by prosecution to support its hypothesis was the examination of two features of the document, the position of some words and the frequency of some letters. In the 1894 trial Alphonse Bertillon acted as an expert witness for prosecution presenting a statistical argument, based upon totally arbitrary assumptions, to show that the probability of chance occurrence in normal handwriting of the above mentioned features was so small that it demonstrated that the document was intentionally written to hide a ciphered message. At the second appeal in 1904, three mathematicians of the French Academy of Science, Jean Gaston Darboux, Paul Émile Appell and Henri Poincaré, were appointed to give their opinion on the reliability of the probabilistic arguments put forward during the first trial. They rejected Bertillon's assumptions, and also expressed an opinion about what should be the correct form of probabilistic reasoning in criminal trials. Given that experts cannot know the *prior probabilities* of the hypotheses at stake, and thus they cannot calculate the *posterior odds*, they must give an opinion only about the value of the *likelihood ratio*:

> Since it is absolutely impossible for us [the experts] to know the a *priori* probability, we cannot say: this coincidence proves that the ratio of the forgery's probability to the inverse probability has a particular value. We can only say: following the observation of this coincidence, this ratio becomes *X* times greater than before the observation.[16]

I call it the rule *Poincaré's rule* because the great scientist knew Bayes' theorem and mentioned it in his 1902 book *The Science and the Hypothesis*. But the rule was forgotten until the end of the century. There are two main reasons that could explain why it happened.

The first reason is that forensic practice was developing in the meanwhile a non-probabilistic *match/no match* approach for giving expert witness in court. The

16 "Examen critique des diverses systèmes ou études graphologiques auxquels a donné lieu le bordereau", in: *L'affaire Dreyfus – La révision du procès de Rennes – enquête de la chambre criminelle de la Cour de Cassation*. Paris: Ligue Française des droits de l'homme et du citoyen 1908, pp. 499-600, p. 504 (English translation by Taroni, Champod and Margot, "Forerunners of Bayesianism in Early Forensic Science", *op. cit*., p. 190). The Odds form Bayes' Theorem says: posterior odds = likelihood ratio x prior odds.

paradigm is fingerprint examination with its rules for identification: an arbitrary threshold is fixed, justified at best by models based upon questionable *a priori* assumptions, or based upon meagre empirical data. Above that threshold, the expert witness is allowed to give a judgment of identification or "compatibility", and below it she will refuse to give witness, or, depending on the case and the rules of evidence of the legal system, she is allowed to grade the answer ("it is possible", etc.).

The second reason has to do with the history of statistics: Fisher's theory of significance tests, and Neyman-Pearson's theory of inductive behaviour came to establish non-Bayesian theories of statistical inference as the dominant paradigm during the 20th century. Both theories do not allow the statistician to assess the relative support given by evidence to competing hypotheses, but ask her to reject or to (provisionally) accept a particular hypothesis. In doing that, these theories fit well with the mainstream *match/no match* approach in forensic practice. Experts trained in classical statistical inference methods simply could not be tuned with the idea that the value of the likelihood ratio determines how evidence discriminates between competing hypotheses. This idea came again to the forefront with the 'Bayesian *renaissance*' in statistical thinking during the second half of the last century. Forensic evidence is a border territory where *classical* and *Bayesian* methods of statistical inference can lead to different, and contradictory, conclusions. A striking example is offered by the debate about *database search*, where the two methodologies dictate contradictory conclusions concerning the value of a match found as a result of a search.

Suppose a search has been made of a data base which contains *DNA* profiles and one of the profiles in the data base matches that of the *DNA* found at the crime scene, and all the other $(N - 1)$ profiles do not match. According to the Bayesian line of argument, the result of the search is evidence for the suspect's guilt and, the larger the database, stronger the evidence. On the contrary, the classical line of argument, supported also by the *U.S. National Research Council* in its 1996 report on *DNA* evidence, claims that, the larger the database, the weaker the evidence against the suspect.[17]

The controversy about database search focuses on the scientific and technological revolution in the field of personal identification that has occurred in 1980s: the emergence of *DNA* typing techniques. This "revolution" has set up a major challenge for forensic practice, opening the door to the use of advanced statistical methods and putting in question the scientific status of traditional forensic practices.

17 See Dawid, "Bayes's Theorem and Weighing Evidence by Juries", in: Richard Swinburne (Ed.), *Ibid.*, pp. 86-87.

4. Is the 'Science of Identification' a scientific practice?

The basic problem for fingerprint analysis is that it is difficult, or even impossible, to calculate observational error rates and frequencies in the relevant populations, and the same is true for all "first generation" forensic methods.[18] With "*DNA* fingerprints" we do have good theoretical models, borrowed from population genetics, to assess the rarity of *DNA* profiles, and the use of standardized technologies would allow to quantify measurement errors and make possible to estimate error rates of laboratories through proficiency tests.[19]

The issue of measurement errors was the *casus belli* of the so called "*DNA* wars" during the 1990s. *DNA* typing was used as evidence in court for the first time in 1987 in U.S.A., and in 1989 the Castro case burst out in New York: Joseph Castro was accused of murdering and prosecution sought to prove that a small bloodstain found on Castro's wristwatch came from the victim. A commercial *DNA* laboratory, *Lifecodes Corporation*, carried out the test and testified that there was a match with the victim's blood. The scientist Eric Lander of *MIT* testified that *Lifecodes*' procedures for interpreting *DNA* results were "so far below reasonable scientific practice in molecular biology as to be appalling".[20] The Castro case started a controversy, involving professionals and academics, about the reliability of *DNA* forensic evidence that offered a valuable contribution to resolve some of the theoretical disagreements and produced two official reports by the *U.S. National Research Council*.[21]

---

18 The term is due to Erin Murphy, "The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence", in: *California Law Review* 95, 2007, pp. 721-797. Other first generation methods are bite and tool mark, hair and fiber, ballistic, handwriting and voiceprint analysis, whereas "second generation" are techniques such as *DNA* typing, biometric scanning, *GPS* and cell-site tracking, *RFID* monitoring and electronic evidence retrieval.

19 John Buckleton, Chris Triggs and Simon Walsh (Eds.), *Forensic DNA Evidence Interpretation*. Boca Raton: CRC Press 2006.

20 Robertson and Vignaux, *Ibid.*, p. 168.

21 Eric Lander, "DNA Fingerprinting on Trial", in: *Nature* 339, 1989, pp. 501-505; Thompson, "Evaluating the Admissibility of New Genetic Identification Tests: Lessons from the DNA War", in*: Journal of Criminal Law and Criminology* 84, 1993, pp. 22-104; Jonathan Koehler, "Error and Exaggerating in the Presentation of DNA Evidence at Trial", in: *Jurimetrics Journal* 34, 1993, pp. 21-39; David Balding and Peter Donnelly, "How Convincing is DNA Evidence?" in: *Nature* 368, 1994, pp. 285-286; Eric Lander and Bruce Budowle, "DNA Fingerprinting Dispute Laid to Rest", in: *Nature* 371, 1994, pp. 735-738; Kathryn Roeder, "DNA Fingerprinting: A Review of the Controversy (with Discussion)", in: *Statistical Science* 9, 1994, pp. 222-278; Balding and Donnelly, "Inference in Forensic Identification (with Discussion)" in: *Journal of the Royal Statistical Society* A, 1995, pp. 21-53; National Research Council, *DNA Technology in Forensic Science*. Washington DC: National Academies Press 1992; National Research Council, *The Evaluation of Forensic DNA Evidence*. Washington DC: National Academies Press 1996. For an overview, see: Jay Aronson, *Genetic Wit-*

The question of error rate estimation is an open, and delicate, question: private and governmental laboratories may carry out proficiency tests and some of them do it on regular basis, but commercial laboratories consider their results as proprietary secrets to be made public only upon lawful request, and governmental agencies even argue that error rates should not be computed: a group of *FBI* leading scientists have recently stated that presenting at trial a specific error rate "adds little value on the discussion on reliability".[22] Thus, on this important point we have only random hints and occasional surveys made by academic scientists.[23]

In scholarly literature a passionate controversy is going on about the scientific validation of fingerprints evidence between scientists and academic scholars on one side and practitioners on the other. Academics maintain that the reliability of latent print individualization is not demonstrated, because the methodology is not scientifically valid, and measurements of the accuracy and validation studies are lacking, so they conclude that latent print examiners' claim to be able to identify with certainty the source of a latent print of unknown origin is unwarranted.[24] Fingerprints professionals reply that the criticism is raised by people who don't have any practical experience in the field, that reliability of latent print identification is demonstrated by the longstanding use of the technique and by the uniqueness of friction ridge skin, that the error rate can be parsed into methodological and human categories and that the methodological error is zero.[25]

---

ness: Science, Law and Controversy in the Making of DNA Profiling. New Brunswick: Rutgers University Press 2007.

22 Bruce Budowle, Maureen Bottrell, Stephen Bunch, Robert Fram, Diana Harrison, Stephen Meagher, Cary Oien, Peter Peterson, Danielle Seiger, Michael Smith, Melissa Smrz, Greg Soltis and Robert Stacey, "A Perspective on Errors, Bias, and Interpretation in the Forensic Sciences and Direction for Continuing Advancement", in: *Journal of Forensic Sciences* 54, 2009, pp. 798-809, p. 801.

23 Around mid-90s, the overall rate of false positive was between 1 in 100 and 1 in 1000 according to Jonathan Koehler, Audrey Chia and Samuel Lindsey, "The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial?", in: *Jurimetrics Journal* 35, 1995, pp. 201-219. The *Cellmark Diagnostics*, which carried out the tests in the O. J. Simpson case, said that its false positive rate was 1 in 200 (Koehler, "One in Millions, Billions, and Trillions: Lessons from People v. Collins (1968) for People v. Simpson (1995)", in: *Journal of Legal Education* 47, 1997, pp. 214-223).

24 Zabell, "Fingerprint Evidence", in: *Journal of Law and Policy* 13, 2005, pp. 143-179. On the issue of validation of fingerprint methodology see Lyn Haber and Ralph Haber, "Scientific Validation of Fingerprint Evidence under Daubert (with Discussion)", in: *Law, Probability and Risk* 7, 2008, pp. 87-150. A very few attempts have been made to build a suitable statistical model, and none of them has been subjected to extended empirical validation studies (see Stoney, "Measurement of Fingerprint Individuality", in: Lee and Gaensslen (Eds.), *op. cit.*). A Bayesian approach attempts to estimate random-match probabilities for fingerprints in the relevant population: Champod and Evett, "A Probabilistic Approach to Fingerprint Evidence", in: *Journal of Forensic Identification* 51, 2001, pp. 101-122.

25 For a list of the relevant literature on the controversy, see: Cole, "Who Speaks for Sci-

In 2009 the *U.S. National Academy of Sciences* has published a two-years study by a panel of experts appointed to investigate non-*DNA* forensic science techniques. The report embraces the scientific perspective, stating that "in a number of forensic science disciplines professionals have yet to establish either the validity of their approach or the accuracy of their conclusion".[26] The report takes a clear stance about the individualization issue: "with the exception of nuclear *DNA* analysis, however, no forensic method has been rigorously shown to be able to individualize", and it states that "the estimation of error rates from proficiency tests are key components of the mission of forensic science", complaining of the "paucity of research in forensic science in support of its claims".[27]

The reaction of the community of professional has consisted mainly in traditional arguments and some *petitio principii*: for example, the *Scientific Working Group on Friction Ridge Analysis, Study, and Technology* (SWGFAST) has affirmed that "history, practice, research have shown that fingerprints can, with a very high degree of certainty, exclude incorrect sources and associate the correct individual to an unknown impression".[28]

The *National Academy* report maintains that claims to identification are unwarranted for all discipline making an exception for *DNA* profiling: "no forensic method other than nuclear DNA analysis has been rigorously shown to have the capacity to consistently and with high degree of certainty support conclusions about individualization".[29] The current *FBI* official policy is to permit examiners to testify that there is an exact match, and so that a person has been identified with

---

ence? A Rto the National Academy of Science Report on Forensic Science", in: *Law, Probability and Risk* 9, 2010, pp. 25-46. The uniqueness argument, even if uniqueness were empirically proved, and it is not, is a *non sequitur*: see Cole, "Forensics without Uniqueness, Conclusion without Individualization: The New Epistemology of Forensic Identification", *op. cit*.; Saks and Koehler, "The Individualization Fallacy in Forensic Science Evidence", in: *Vanderbilt Law Review* 61, 2008, pp. 199- 219.

26 National Research Council, Committee on identifying the needs of the forensic science community, *Strengthening Forensic Science in the United States: A Path Forward*, Washington DC: National Academies Press 2009, p. 53.

27 National Research Council, *op. cit.*, pp. 7, 122, and 186. "The problems that the report identifies – which include inconsistent requirements for training, certification and accreditation, insufficiently rigorous protocols, lack of oversight, inadequate testing and validation of elementary principles, unaided subjective interpretation of data, poor acknowledgment, understanding and measurement of potential sources of biases and errors and so forth – will not disappear on their own when funding for the forensic sciences increases. As the report suggests, substantial progress can only be made if there is a genuine commitment to scientific reform within the forensic science community and structural support outside of it. Unfortunately, there are reasons to be skeptical about the will of the requisite communities." (Koehler, "Forensic Science Reform in the 21st Century: A Major Conference, a Blockbuster Report and Reasons to be Pessimistic", in: *Law, Probability and Risk* 9, 2010, pp. 1-6, especially, p. 3).

28 *SWGFAST NAS Position Statement*, 8/3/2009. Posted at *http://www.swgfast.org/*.

29 National Research Council, *op. cit.*, p. 87.

certainty, when the likelihood of a random match is less than 1 in 260 billion. It is a very high probability, but the truth of a statement of identification does not logically follow from it, and in doing that statement, the examiner is jumping to a conclusion about a factual hypothesis: "The movement from a probability statement to one of certainty represents a 'leap of faith' rather than a logical consequence".[30]

The probabilistic nature of evidential reasoning represents a strong cognitive burden for common sense thinking, especially in the field of law, where consequences of decisions are of great importance, and the *match/no-match* approach in forensic practice has been a way to avoid such a burden. The application of good techniques of rational reasoning would be a major step forward in the "scientification of forensic practice". Some authors claim that recognizing that the process of identifying an individual as being the source of a crime mark is a *decision process*, and that applying the techniques of Bayesian decision theory would greatly help in promoting accurate judicial decision making.[31] In this perspective the task of the forensic scientist is to assist the probabilistic evaluation of evidence by presenting her findings through the appropriate use of Poincaré's rule, and the very concept of individualization may be disposed of.[32] The "scientification of forensic practice" turns out to be another aspect of the overall important issue of the best use of science and technology in our society.

Faculty of Design and Arts
IUAV University
Dorsoduro, 2206
30123, Venice
Italy
pgarboli@iuav.it

30 Aitken and Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, *op. cit.*, pp. 86-87. The term "leap of faith" was coined by Stoney, "What Made us ever Think we Could Individualize Using Statistics?", *op. cit*.

31 Alex Biedermann, Silvia Bozza, Franco Taroni, "Decision Theoretic Properties of Forensic Identification: Underlying Logic and Argumentative Implications", in: *Forensic Science International* 177, 2008, pp. 120-132; Taroni, Bozza, Biedermann, Garbolino and Aitken, *Data Analysis in Forensic Science: A Bayesian Decision Perspective*. Chichester: Wiley 2010.

32 Some authors have even made a call for a "paradigm shift" in forensic science: Saks and Koehler, "The Coming Paradigm Shift in Forensic Identification Science", in: *Science* 309, 2005, pp. 892-895.

Wenceslao J. Gonzalez

The Sciences of Design as Sciences of Complexity: The Dynamic Trait[1]

Abstract

The sciences of design can be analyzed as sciences of complexity. This involves taking into account the twofold complexity in science: the structural and the dynamic. Thus, the analysis can move from structural complexity to dynamic complexity. Here the focus is on the dynamic trait, which means the study of change in complex dynamics. In this regard, there are three main notions: process, evolution, and historicity. This paper draws attention to the need for historicity in human-made disciplines which include the emphasis on "activity" rather than on "behavior".

1. Twofold complexity in science: Structural and dynamic

A central feature of many sciences is *complexity*,[2] which affects problems, methods and results of scientific research. This characteristic is twofold insofar as complexity concerns the structure and the dynamics of a given science or a set of sciences.[3] Thus, complexity can be focused either from the structural perspective or from the dynamic viewpoint. In the first case, the study of complexity is with regard to the framework or constitutive elements present in a science or group of sciences, whereas in the second possibility the analysis of complexity is related to change over time of the motley elements involved in that science or collection of sciences, taking into account the forces generating the change.[4]

1 This research project is supported by the Spanish Ministry of Science and Innovation (FFI2008-05948).

2 Complexity is a topic that might be focused from quite different angles, cf. Klaus Mainzer, *Thinking in Complexity*. *The Computational Dynamics of Matter, Mind, and Mankind*. Berlin: Springer 2007, 5th ed.

3 In the case of economics, which is the discipline central to this paper, this can be seen in the papers collected in the three volumes of John Barkley Rosser Jr (Ed.), *Complexity in Economics*. Cheltenham: E. Elgar 2004.

4 These categories of structural and dynamic can be used to articulate lists of kinds of complexity such as "multilevel organization, multicomponent causal interactions, plasticity in relation to context variation, and evolved contingency", Sandra D. Mitchell, *Unsimple Truth: Science, Complexity, and Policy*. Chicago: The University of Chicago Press 2009, p. 21.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_24,

When the point of view of the analysis is dynamic, the emphasis is on notions such as "process", "evolution" or "historicity." Hence, the scientific attention is on transitions, variations or modifications between a previous stage and a posterior one. Thus, in the scientific context of the examination of complexity from a dynamic viewpoint, *dynamics* is not merely a type of attribute that has a relation to a possible potential power ("dynamis"). In principle, dynamics should also be connected to something with actuality or actively existing ("energeia") that involves change over time.

These complex aspects of the world that change in different ways might be natural, social, or artificial. Thus, the variations concern the three groups of sciences. Consequently, this kind of change, which is analyzed scientifically in terms of dynamic complexity, is also present in the sciences of design, such as economics.[5] *De facto*, the complex dynamics of economics receives frequent attention, mainly in the sphere of macroeconomics (e.g., market mechanisms, business cycles, economic growth, economic development, etc.),[6] where there are commonly more factors involved than in the realm of microeconomics.

As a matter of fact, the sciences of design have *twofold complexity*: a complexity in their constitutive components – the complex framework – and a complexity in the dynamics, which involves aims, processes, and results. In these disciplines, the ingredients of the complex dynamics might be seen when the scientific elements operate as a teleological procedure open to many possibilities in the future. Moreover, insofar as the sciences of design are developed as applied sciences, as happens in economics, there is a combination of prediction and prescription.[7] This feature towards the future increases the dynamic complexity of this science.

Here the main interest is in the *dynamic trait* of the sciences of design understood as sciences of complexity. Thereafter, the attention shifts to the repercussion of the dynamic complexity for making economic predictions. Thus, this paper complements an earlier one on "Complexity in Economics and Prediction: The Role of Parsimonious Factors",[8] which was primordially oriented towards struc-

---

5 Cf. Herbert Simon, *The Sciences of the Artificial*. 3rd ed., Cambridge (Mass.): The MIT Press 1996; and Simon, "Organizing and Coordinating Talk and Silence in Organizations", in: *Industrial and Corporate Change* 11, 3, 2002, pp. 611-618.

6 A good example can be found in Richard Day, *Complex Economic Dynamics*. Vol. I, Cambridge (Mass.): The MIT Press 1994; and Day, *Complex Economic Dynamics*. Vol. II, Cambridge (Mass.): The MIT Press 1999. The first volume focuses on dynamical systems and market mechanisms, whereas the second volume analyzes macroeconomic dynamics.

7 On the general angle, see Simon, "Prediction and Prescription in Systems Modeling", in: *Operations Research* 38, 1990, pp. 7-14. On the specific economic case, see Wenceslao J. Gonzalez, "Prediction and Prescription in Economics: A Philosophical and Methodological Approach", in: *Theoria* 13, 32, 1998, pp. 321-345.

8 Cf. Gonzalez, "Complexity in Economics and Prediction: The Role of Parsimonious Factors", in: Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Thomas Uebel, and Marcel Weber (Eds.), *Explanation, Prediction, and Confirmation*. Dordrecht:

tural complexity. In this paper, the next steps are the following: (i) the relation between structural complexity and dynamic complexity; (ii) the change in complex dynamics in terms of process, evolution and historicity; and (iii) the need for historicity in human-made disciplines, taking into account the differences between behavior and activity.

2. From structural complexity to dynamic complexity

On the one hand, in the configuration of a science of design there is a structural complexity. This *complex framework* can be seen in the constitutive elements of a science of the artificial, such as language, structure, knowledge, method, activity, aims, and values. They can be analyzed in economics as a science in the realm of the human-made. On the other hand, a science of design involves a *complexity in the dynamics*, especially when it is working as an applied science. So, a science of design is a teleological human activity that seeks the solution of concrete problems. It uses a complex practical system organized by aims, processes, and results. These are very noticeable in applied economics when dealing with problems such as the ongoing economic financial crises.

Initially, this twofold complexity, which is available in a science of design like economics, leads to two general philosophic-methodological issues: a) how the structural complexity articulates with the dynamic complexity (in this case, as a human-made discipline); and b) how the second evolves over time, introducing changes, including the revolutionary modifications, and this requires one to consider what the adequate concepts to grasp the changes in a complex dynamics are. These issues are particularly relevant for economic predictions, because they are commonly related to dynamic systems and need to deal with important changes, both in the "internal" side of economics ("economic activity") and in the "external" environment around economics ("economics as activity").

Regarding the articulation between both kinds of complexity, it seems clear that "structural" is interwoven with "dynamical" when a structure has a function or when it is a means for an end. In addition, there is the configuration of the complex structure, which requires some dynamic procedures,[9] and a relation between a multifaceted whole and its parts is not static in many cases. Nicholas Rescher includes several of these aspects in his characterization of complexity,[10]

Springer 2011, pp. 319-330.

9 According to John Foster and Stan J. Metcalfe, in the case of economics, "complex systems are network structures and should be dealt with as networks, not collapsed into analytical functional relationships, such as the production functions that underpin most of conventional growth models", Foster and Metcalfe, "Evolution and Economic Complexity: An Overview", in: *Economics of Innovation and New Technology* 18, 7, 2009, p. 609.

10 Cf. Nicholas Rescher, *Complexity*: *A Philosophical Overview*. New Brunswick, NJ:

even though his approach is mainly related to "structural complexity" rather than focusing on "dynamic complexity." He distinguishes *epistemic modes* of complexity (descriptive, generative, and computational) and *ontological modes* of complexity (compositional, structural – i.e., organizational and hierarchical –, and functional complexity).[11]

Rescher's characterization is open to some *dynamic aspects* of complexity that are relevant for a science of design. These dynamic aspects might be seen in the generative complexity, within the epistemic modes of complexity, as well as in the operational complexity and the nomic complexity, which belong to the group of ontological modes. Following this view, the complex structure of a science of design, such as economics, requires a development over time, either to produce the complex system at stake ("generative complexity") or to make the variety of types of functioning ("operational complexity") and the possible laws or norms governing the phenomena at issue ("nomic complexity").

According to this characterization of complexity, oriented towards epistemology and ontology, Rescher criticizes Simon explicitly for being too general in his conception of "complexity." This author sees a complex system mainly in structural terms based on a hololological perspective. Thus, for Simon, a complex system is "one made up of a large number of parts that have many interactions."[12] In such systems, "given the properties of the parts and the laws of their interaction, it is not trivial matter to infer the properties of the whole."[13] Meanwhile, for Rescher, this description is not particularly helpful, since few things (natural, social, or artificial) seem exempt from this organizational principle.[14] He considers that the emphasis on complexity cannot be in "the extent to which chance, randomness, and the lack of lawful regularity in general is absent."[15]

Obviously, when the system follows laws or norms, these rules can be more or less complex. Furthermore, the changes over time might be somehow shallow ("evolutionary") or clearly deep ("revolutionary") as well as continuous or discontinuous. In his view, Rescher highlights that complexity might be functional and, therefore, it can be complex in dynamic terms. Thus, when the systems are goal-directed in their *modus operandi*, as is commonly the case in the sciences of design,[16] they do this "generally towards a plurality of potentially competing

---

Transaction Publishers 1998, pp. 1-26; especially, pp. 8-16.

11 Cf. Rescher, *Complexity*: *A Philosophical Overview*, p. 9.

12 Simon, *The Sciences of the Artificial*. 3rd ed., pp. 183-184.

13 Simon, *Ibid.*, p. 184.

14 Cf. Rescher, *Complexity*: *A Philosophical Overview*, p. 22, note 14.

15 Rescher, *Ibid*, p. 8.

16 Cf. Gonzalez, "Rationality and Prediction in the Sciences of the Artificial: Economics as a Design Science", in: Maria Carla Galavotti, Roberto Scazzieri, and Patrick Suppes (Eds.), *Reasoning, Rationality and Probability*. Stanford, CA: CSLI Publications 2008, pp. 165-186; especially, pp. 169-171.

goals."[17] He thinks that there might be in the system a complexity *operational*, which is "displaying dynamic complexity in the temporal unfolding of its processes", or it can be *nomic*, which is "a timeless complexity in the working interrelationships of its elements."[18]

Certainly, Simon is not unaware of the existence of a complex dynamics, insofar as he "explores the dynamic properties of hierarchically organized systems and shows how they can be decomposed into subsystems in order to analyze their behavior."[19] But his view of dynamic complexity in the science of design is restricted, insofar as he is primarily oriented to the evolution of complex systems that are usually hierarchical. For him, "among possible forms, hierarchies are the ones that have time to evolve."[20] In addition, he thinks that there are systems where "the whole is more than the sum of the parts",[21] which seems open to the idea of emergent properties.[22] This is the case in economics, where the complexity of economic structures can lead to emergent properties.[23]

Meanwhile Rescher has, in principle, a wider scheme of things regarding dynamic complexity in science than Simon, due to his pragmatic approach connected to a metaphysical realism.[24] When Rescher analyzes complex changes in science, he is commonly thinking in terms of *processes*, whereas Simon tends to think of complexity as an adaptation of an evolutionary kind. But it seems to me that both approaches are insufficient to grasp actual dynamic complexity of a science of design. Thus, besides the valuable set of possibilities that Simon and Rescher have presented, we need to think of enlarging the collection of options of dynamic complexity in order to complete the main elements of complexity.

J. Barkley Rosser takes that direction and offers us a broader vision of the complexities of complex dynamics.[25] His view is open to a "historicity" in the analysis of the dynamics of complex systems, insofar as he emphasizes the exist-

17 Rescher, *Ibid*, p. 15.

18 Rescher, *Ibid*, p. 12.

19 Simon, *The Sciences of the Artificial*. 3rd ed., p. 184. Cf. Simon, "Near Decomposability and the Speed of Evolution", in: *Industrial and Corporate Change* 11, 3, 2002, pp. 587-599.

20 *The Sciences of the Artificial*. 3rd ed., p. 197. See also in the same book pages 188-190.

21 Simon, *Ibid*, p. 184.

22 "The prospects for the emergence of an effective complex system are much greater if it has a nearly-decomposable architecture", Simon, "Complex Systems: The Interplay of Organizations and Markets in Contemporary Society", in: *Computational and Mathematical Organizational Theory* 7, 2001, p. 82.

23 Cf. Karl-Ernst Schenk, "Complexity of Economic Structures and Emergent Properties", in: *Journal of Evolutionary Economics* 16, 2006, pp. 231-253.

24 Cf. Rescher, "Pragmatic Idealism and Metaphysical Realism", in: John R. Shook and Joseph Margolis (Eds.), *A Companion to Pragmatism*. Oxford: Blackwell 2006, pp. 386-397.

25 Cf. Barkley Rosser Jr, "On the Complexities of Complex Economic Dynamics", in: *Journal of Economic Perspectives* 13, 4, 1999, pp. 169-192.

ence of discontinuities in their changes (including catastrophes). He maintains that the studies of complexity in a variety of disciplines, including economics, have evolved out of earlier work using nonlinear dynamics. They have been used to explain such phenomena as path dependence in technological evolution and regional development as well as "the appearance of discontinuities, such as the crashes of speculative bubbles or the collapses of whole economic systems." [26]

2.1 Process, evolution, and historicity: The change in complex dynamics

Besides the articulation of structural complexity and dynamic complexity, the issue of how the complex dynamics changes over time is crucial. In this regard, complex dynamics can be characterized at least in three terms: process, evolution and historicity. The first – process – is quite general, but it is certainly needed for the analysis of complexity in a science such as economics.[27] The term has a metaphysical basis, as Rescher has pointed out in his criticism of Peter Strawson.[28] The second term – evolution – is extremely frequent when dealing with dynamic complexity,[29] and certainly "evolution" is not incorrect in this realm. But it seems to me that evolution is insufficient to cover the whole field of dynamic complexity related to the sciences of design, in general, or economic predictions, in particular. The third term – "historicity" – seems more in tune with the reality of complex changes in economics. Moreover, historicity might be considered as a key factor for understanding the problems for economic predictions.

"Process" is a term explicitly assumed by Richard Day when he maintains that "*complex dynamics* include processes that involve nonperiodic fluctuations, overlapping waves, switches in regime (…). These types of change are very different than the stationary states, periodic cycles, and balanced paths of growth. But they are ubiquitous phenomena in the economics of experience."[30] In this regard, concerning complex dynamics in general, Rescher is particularly keen on the notion of *process*.[31] His perspective seems useful for contextual aspects of economics (e.g., those related to technological innovations), because the key distinction

26 Barkley Rosser Jr, *Ibid*, p. 169.

27 A proposal in this regard, focused on economics, can be found in Martin Shubik and Eric Smith, "Building Theories of Economic Process", in: *Complexity* 14, 3, 2008, pp. 77-92.

28 Cf. Rescher, *Process Methaphysics*. Albany: State University N. York Press 1995, pp. 60-62.

29 A classical example is Philip W. Anderson, Keneth J. Arrow, and David Pines (Eds.), *The Economy as an Evolving Complex System*, Santa Fe Institute, Santa Fe, NM, 1988.

30 Day, *Complex Economic Dynamics*. Vol. I, p. 4.

31 Rescher has also developed a set of ideas regarding evolution, cf. Rescher, *A Useful Inheritance. Evolutionary Aspects of the Theory of Knowledge*. Savage, MD: Rowman and Littlefield 1990. But "process" seems a more basic notion insofar as he discusses the "Varieties of Evolutionary Process", pp. 5-12.

is – for him – between "product-productive processes" and "state-transformative processes."

The first type is the process that produces what can be characterized as something tangible, a thing or "substance" (e.g., the manufacturing processes that produces a medicine); and the second type is the process that merely transform states of affairs, paving the way for further processes without issuing in particular things or states thereof (e.g., windstorms).[32] In addition, Rescher emphasized the existence of *owned* and *unowned* processes, where the former is connected with agents (which is frequently the case of economics), whereas the latter does not represent the activity of actual agents (e.g., the fluctuation of a magnetic field).[33]

Another term closely related to the change in complex dynamics is "evolution."[34] Moreover, the idea of evolution is particularly frequent in the analysis of dynamic complexity in science, including sciences of design such as economics. Thus, it happens that, among the economists interested in complexity, a number of influential authors recognize the evolutionary influence – either large or small – on their views, such as Friedrich Hayek, Joseph Schumpeter, Herbert Simon, Reinhard Selten, etc.[35] Undoubtedly, *evolution* is a term that can be understood in quite different ways, but the evolutionary approach to economics is *de facto* dominated by schemes that are mainly Lamarckian or Darwinian.[36]

Commonly, when dynamic complexity in economics is related to evolutionary concepts, this is due to the confluence of three different strands of analysis: a) the early biological grafting, where the economic system was understood as an organism with important similarities to biological entities; b) the conception of learning as an "engine of growth", because the generation of new knowledge leads to innovation in business firms; and c) the relation between rationality and change – the behavioral theory of economics –, where Simon has a key role with the notion of "bounded rationality" and the distinction between "substantive" and "procedural" rationality.[37]

Economics, as whole, appears then as a huge evolutionary system. Its dynamics as "an organism" includes knowledge as a key element, understood as a source

32 Cf. Rescher, *Process Methaphysics*, p. 41.

33 Cf. Rescher, *Ibid.*, p. 42.

34 Cf. Lawrence E. Blume, and Steven N. Durlauf (Eds.), *The Economy as an Evolving Complex System, III: Current Perspectives and Future Directions*. New York: Oxford University Press 2006.

35 Cf. Gonzalez, "Evolutionism from a Contemporary Viewpoint: The Philosophical-Methodological Approach", in: Gonzalez (Ed.), *Evolutionism: Present Approaches*. A Coruña: Netbiblo 2008, pp. 40-41.

36 Cf. Geoffrey M. Hodgson, "Is Social Evolution Lamarckian or Darwinian?", in: John Laurent and John Nightingale (Eds.), *Darwinian and Evolutionary Economics*. Cheltenham: E. Elgar 2001, pp. 87-120; especially, p. 88.

37 An analysis of these elements is in Cristiano Antonelli, "The Economics of Innovation: From the Classical Legacies to the Economics of Complexity", in: *Economics of Innovation and New Technology* 18, 7, 2009, pp. 611–646; especially, pp. 629-633.

of innovation. Furthermore, this huge system involves a behavior that is context–dependent (it might frequently have plasticity in the variation). In this general view on evolution of a system, there is a connection between complex dynamics and prediction as a key methodological ingredient of economics. In this regard, for Bertuglia and Vaio, complexity

> does not mean a confused forecasting: it simply means that it is impossible to build up a model which can account for the sudden and (most of all) unexpected 'changes' that sometimes take place during the evolution of a system, even though the evolutionary path between one change and the next can be well described by deterministic laws.[38]

Changes in Simon are evolutionary insofar as the mechanisms used are adaptive. He insisted on the cognitive contents of the economic changes – the agents making decisions with bounded rationality – and he used to analyze these changes in terms of an evolution that is supported by the adaptive rationality of the agents.[39] Explicitly, he recommended the analysis made by Nelson and Winter in the book *An Evolutionary Theory of Economic Change*.[40] Usually, Simon was more interested in the architecture of complexity than in the dynamics of complex systems.[41] His priority was commonly in how the sciences of complex systems can be configured,[42] even though he also paid attention to some aspects of the evolution of these systems.[43] In this regard, his focus normally was on a rationality adapted to the context and in the "dynamic properties of hierarchically structured systems."[44]

However, besides the notion of "process" and the idea of "evolution", there are other options on the change of dynamic complexity. The third main option is the emphasis on the concept of "historicity." This perspective involves a deeper analysis of the complex change over time of something artificial. This should be made in order to grasp the variations in the realm of the sciences of design, in general, and in economics in particular. Moreover, if the aim is to tackle problems

38 Cristoforo S. Bertuglia and Franco Vaio, *Nonlinearity, Chaos and Complexity. The Dynamics of Natural and Social Systems*. Oxford: Oxford University Press 2005, p. vii.

39 Cf. Simon, *Reason in Human Affairs*. Stanford, CA: Stanford University Press 1983. This does not mean "a passive" attitude regarding the future, cf. Simon, "Forecasting the Future or Shaping it?", in: *Industrial and Corporate Change* 11, 3, 2002, pp. 601-605.

40 Richard N. Nelson and Sidney G. Winter, *An Evolutionary Theory of Economic Change*. Cambridge (Mass.): Harvard University Press 1982. I am among those who received this recommendation.

41 Cf. Simon, "The Architecture of Complexity", in: *Proceedings of the American Philosophical Society* 106, 6, 1962, pp. 467-482.

42 Cf. Simon, "Can There Be a Science of Complex Systems?", in: Yaneer Bar-Yam (Ed.), *Unifying Themes in Complex Systems: Proceedings from the International Conference on Complex Systems 1997*. Cambridge (Mass.): Perseus Press 1999, pp. 4-14.

43 Cf. Simon, "Near Decomposability and the Speed of Evolution", pp. 587-599.

44 Simon, *The Sciences of the Artificial*. 3rd ed., p. 184.

such as reliability of economic predictions, the analysis on dynamic complexity should take into account "historicity", which in principle goes beyond "process" and "evolution."

Concerning economics as a historical science, Simon wrote a paper.[45] In this article he "examines some of the ways in which history and economics can be fashioned into economic history."[46] His main reasons for this link were related to human agents and their changes in representations, perceptions and motivations, due to their interaction with the environment. According to Simon's viewpoint, 1) from time to time, boundedly rational economic actors represent the economic scene in radically different ways; and 2) these changes in representation are connected to two aspects: a) the changes occur as a function of natural and social events, social influences on perception; and b) there are variations on the molding of human motives by the social environment, which is itself time dependent.[47]

Due to these and other reasons, which Simon sees as "bound closely to basic human characteristics, the dynamic movements of the economic system depend not only on invariant laws, but on continually changing boundary conditions as well."[48] But his vision of history is here rather external, devoted mainly to contextual aspects, instead of being also genuinely internal, involved with legitimate scientific contents. Thus, he wrote that "science deals with invariants and history with dated events."[49] Moreover, his acceptance of the historical approach in economics is merely based on a similitude with the natural sciences, because they have differential equations filled with time derivatives.[50] In this regard, Barkley Rosser has pointed out the existence of aspects of complexity in economics that makes its dynamics different from physics. The additional layer of complexity comes from the interaction of human calculations in decision-making.[51] Certainly, this is the case of a science of design such as economics (e.g., in urban planning).

45 Cf. Simon, "Economics as a Historical Science", in: *Theoria* 13, 32, 1998, pp. 241-260.

46 Simon, *Ibid*., p. 241.

47 Cf. *Ibid*., p. 241. See also Simon, "Bounded Rationality in Social Science: Today and Tomorrow", in: *Mind and Society*, 1, 1, 2000, pp. 25-39.

48 Simon, "Economics as a Historical Science", p. 241.

49 Simon, *Ibid.*, p. 241.

50 "The prevalence in the natural sciences and economics of differential equations filled with time derivatives should persuade us of the legitimacy of joining history with science", Simon, *Ibid*., p. 241.

51 "Although complexity is a multidisciplinary concept derived from mathematics and physics, the extra complications arising in economics because of the problem of interacting human calculations in decision-making add a layer of complexity that may not exist in other disciplines", Barkley Rosser Jr, "On the Complexities of Complex Economic Dynamics", p. 171.

2.2 The need for historicity in human-made disciplines: From behavior to activity

Any human-made discipline, as is the case of economics as a science of design, is historical, and historicity is a key factor for its dynamic complexity. On the one hand, there are certainly economic events to be dated according to accepted chronological criteria, and, on the other, as human-made undertakings, economic events are *eo ipso* historical insofar as they born with the feature of revocability according to internal and external criteria to the endeavor itself. Thus, factors such as originality in the theories suggested, creativity in the designs proposed, innovation in the technology used and the like are only possible in economics on the basis of a historicity of the endeavor developed.[52]

(i) Dynamic complexity in economics initially has its roots in the historical aspects that are present in any science. Every science is *our* science, and each discipline has problems, models and results that are historically conditioned, both in internal terms and in external terms. (ii) There is the dynamic complexity that is related to the features of a science of design, insofar as it is an applied science and, therefore, economics has aims, processes and results which are context-dependant. Thus, the economic solutions for ongoing financial problems cannot be the identical to those solutions given many years ago, because the circumstances are not the same. (iii) There is the dynamic complexity that comes from the agents themselves, those who develop a science of design such as economics as operative subjects. *De facto*, the decision-making of the agents is crucial many times in economic matters.

Simon is aware of the existence of problems in dynamic complexity, and this is one of his reasons for the search of parsimonious factors in science.[53] He has made remarks regarding the levels two and three just pointed out. Thus, his approach to economic dynamics considers some contextual elements of the complex dynamics of economics as a science of design, mainly those that might be connected with structural components to be adjusted to a changeable environment. In addition, he pays attention to the historical ingredients in the decision-making of the agents, seeing how the behavior of the human agents uses bounded rationality as an instrumental rationality of an adaptive kind.

For him, the "dynamics in economic history" includes some relevant aspects: a) technological change, b) the institutional context, and c) certain categories of exogenous institutional variables, such as changes in the utility function, in the

52 Cf. Gonzalez (Ed.), *Racionalidad, historicidad y predicción en Herbert A. Simon*. A Coruña: Netbiblo 2003, and Gonzalez (Ed.), *Las Ciencias de Diseño: Racionalidad limitada, predicción y prescripción*. A Coruña: Netbiblo 2007.

53 Cf. Simon, "Science Seeks Parsimony, not Simplicity: Searching for Pattern in Phenomena", in: Arnold Zellner, Hugo A. Keuzenkamp, and Michael McAleer (Eds.), *Simplicity, Inference and Modelling. Keeping it Sophisticatedly Simple*. Cambridge: Cambridge University Press 2001, pp. 32-72.

production function, and in the laws of property.[54] Although Simon insists on exogenous elements, he also accepts an endogenous component: "we usually think of history as a process of continuing change, something to be captured by dynamic models, like differential equation models for predicting business cycles and other movements in economic activity."[55]

Although Simon has been critical with the neoclassical approach – the mainstream economics –, he recognizes that "the variables considered thus far are all consistent with the assumptions of neoclassical theory."[56] This is the case of the exogenous elements pointed out here as well as the use of comparative statistics for historical events (urbanization, railroads, etc.), which he also accepts. His contribution lies then in connecting bounded rationality and economic dynamics.

Thus, Simon proposes adding "the variables that deal with the fact that human rationality is bounded. These additional variables are closely bound with a historical view of economics, for they take into account (1) continuing changes in knowledge and information (both knowledge about economics and other knowledge about the world), (2) changes in human ability to estimate consequences of actions, (3) changes in the institutional setting within which economic behavior takes place, (4) changes in the focus of attention and related changes in beliefs and expectations. I will add, for they belong among the belief-dependent variables, (5) changes in human altruism and (6) in group identification."[57]

Again, Simon is thinking of an *economic behavior* that takes place in a complex natural and social environment, where it should be an evolutionary adaptation of an agent or an organism. Insofar as this environment remains exogenous, the "laws will continue to change with changes in social institutions and changes in the knowledge and beliefs of the boundedly rational people who inhabit them. The focus of individual and public attention will shift with changing events from one set of variables to another, with resulting shifts in individual and system behavior."[58] Creativity, innovation and originality of the agents do not seem to have a relevant space here.

Even though Simon's conception has advantages in comparison with the neoclassical approach, it is less complete than the perspective on economics based on *human activity*. The complexity of economic reality – especially, the complex

54 In his view, "a partial catalogue of exogenous institutional variables (and candidates for endogenization) would include: (1) changes in the utility function, with consequent changes in demand and in savings rates; (2) changes in the production function, resulting from technological change and other factors, and with consequent changes in supply; and (3) changes in the laws of property, with consequent effects upon positive and negative externalities, the appropriability of inventions and powers of government to redistribute income and wealth." Simon, "Economics as a Historical Science", pp. 250-251.

55 Simon, *Ibid*., p. 248.

56 *Ibid*., p. 251.

57 Simon, *Ibid*., p. 251.

58 *Ibid*., p. 258.

dynamics – is better analyzed in terms of the dual components "economic activity" and "economics as activity" than in "economic behavior." Thus, there is an *economic activity*, something which could be understood as autonomous regarding other human activities. It comprises economic activity which human beings carry out in their interrelations involving goods and services, exchanges and commodities, innovation and plan optimizing decisions, and so. Meanwhile, *economics as activity* connects the links between economic activity and other human activities (political, sociological, cultural, ecological, …). In this case, economic activity appears integrated into the whole system of human relations; it is immersed in the set of activities developed by human beings in normal circumstances. Then, as an activity *among* others, economics has links with many activities (political, sociological, cultural, ecological, …).[59]

Both – economic activity and economics as activity – should be considered here, because this science explains and predicts human activities in the domain of a concrete sphere (i. e., exchange, commodities, …). These elements have direct implications for the realm of prediction. On the one hand, the normal aim of a human activity is more connected with present circumstances than with a future not yet observed. On the other hand, the predictability of economic activity – which is, in principle, autonomous – is possible, and could be reliable; whereas predictability of economics as a human activity among others appears more unreliable, due precisely to the interdependence with other activities. Hence, prediction does not appear as the *central aim* of economics, in spite of the predictivist thesis of neoclassical economics,[60] and its scientific character could be accepted in the economic activity.

Between "human activity" and "human behavior" there are several differences. a) Activity has an immediate *practical* character: it includes *praxis* – it is doing something which affects its reality –, whereas behavior has a less diversified scope, mainly when it is understood as instinctive (close to animal behavior). b) Activity has in itself *historicity*: human activity is *eo ipso* historical, not only in the sense of *having* time, but also in the deepest sense of occurring and developing precisely *with* time. This historicity affects the decision making process and it should be included among the elements to be studied. Behavior, on the contrary, has a more static constitution, because it can be considered without especial concern for historicity (a very well known example is behaviorism). c) Activity has a very close link with *language*, more than behavior. So, there is no problem in the connection between action and language (such as in the case of "speech acts") whereas there are criticisms regarding behaviour and language (e.g., with Skinner's "verbal behavior" or Quine's proposals). d) Activity has both a *descriptive*

59 Cf. Gonzalez, "Economic Prediction and Human Activity. An Analysis of Prediction in Economics from Action Theory", in: *Epistemologia* 17, 1994, pp. 253-294.

60 Cf. Gonzalez, "Prediction as Scientific Test of Economics", in: Wenceslao J. Gonzalez, and Jesus Alcolea (Eds.), *Contemporary Perspectives in Philosophy and Methodology of Science*. A Coruña: Netbiblo 2006, pp. 83-112.

and a *normative* sphere, because there are genuine social actions which require norms to rule it properly (either ethically or legally), whereas behavior is more descriptive than normative.[61]

Faculty of Humanities
University of A Coruña
Dr. Vazquez Cabrera street, w/n
15.403, Ferrol
Spain
wencglez@udc.es

61 Cf. Gonzalez, "Racionalidad y Economía: De la racionalidad de la Economía como Ciencia a la racionalidad de los agentes económicos", in: Gonzalez (Ed.), *Racionalidad, historicidad y predicción en Herbert A. Simon*, pp. 88-89.

Subrata Dasgupta

Epistemic Complexity and the Sciences of the Artificial

Abstract

In 1962 Herbert Simon articulated the nature of complexity of both natural and artificial systems. A system, he said, is complex if it is composed of a large number of components that interact in nontrivial ways. I will label Simon's notion as *systemic complexity*. However, in the case of *artifacts* – things produced or conceived in response to some need or desire – there is another type of complexity which is especially relevant. This is the *richness of the knowledge embedded in an artifact*. I call this *epistemic complexity*. It comprises of the knowledge that both contributes to the creation of an artifact and the knowledge generated as a result of that creation.

Insofar as artifacts are what the *sciences of the artificial* are about, we might hope that the study of epistemic complexity might deepen our understanding of the sciences of the artificial and the nature of artifact creation.

In this paper I use examples from the history of technological artifacts to analyze aspects of epistemic complexity and its relation to systemic complexity.

1. Two types of complexity

In 1962, Herbert Simon articulated the nature of complexity as it is evident in both natural and artificial systems. A system, he said, is said to be complex if it is composed of a large number of components that interact in nontrivial ways. This means that even if one understands the properties of each component in isolation, one may not be able to interpret the properties of the system as a whole.[1]

I will label Simon's notion as *systemic complexity*. Now, *artifacts* – objects that are produced or conceived in response to some need or desire – are clearly more or less complex in this systemic sense. But there is another type of complexity which is especially relevant in the case of artifacts. And this is *the richness of the knowledge that is embedded in an artifact.* I will call this *epistemic complexity.* It comprises of the knowledge that both contributes to the creation of an artifact; and the knowledge that is generated as a result of that creation. Insofar as *arti-*

1 Herbert A. Simon, "The Architecture of Complexity", in: *Proceedings of the American Philosophical Society*, 106, 1962, pp. 467-482; Simon, *The Sciences of the Artificial.* Cambridge (Mass.): The MIT Press 1996.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_25,

facts are what the "sciences of the artificial" are about,[2] examination of epistemic complexity contributes, I believe, to our understanding of the nature of artifacts. Insofar as the systematic *study* of the nature of artifacts is what the "sciences of the artificial" are about, I will hope that shedding light on epistemic complexity will contribute to these sciences.

The nature of the complexity of artifacts has been of interest to me for many years reaching back to my study of the structure of design processes in the realm of computer system design and the design of languages to describe such systems.[3] This paper presents, somewhat briefly, some of the results of these studies especially as they relate to the epistemic complexity of artifacts and its relationship to systemic complexity.

Before I continue let me introduce a term of convenience. Henceforth, I will refer to any *practitioner* who creates artifacts as *artificer*. This is a somewhat archaic word but accurate nonetheless. It embraces inventors, designers, engineers, technologists. I will also use the collective term *artifactual creation* to include design, invention and making.

2. Technological knowledge and epistemic complexity

Artifactual creation is a *knowledge rich cognitive process*. The artificer is armed with a rich body of interconnected knowledge and beliefs which he or she brings to bear in any particular cognitive act of creation.[4] Some of this knowledge is shared by people in general, not just artificers, e.g., common rules of inference, or general mental tools for planning and problem solving. More specific artifactual knowledge is itself quite varied. It includes, e.g., mathematics, the basis sciences and engineering theory. But these types of knowledge have entered the artificer's mind relatively recently, mostly since the 18th century and the Industrial Revolution.[5] In the very long history of artifactual creation, reaching back to the origins of humankind itself, the dominant form of knowledge is what, following Michael Polanyi,[6] we may call *operational principles*. This term refers to all rules, proce-

2 Simon, *The Sciences of the Artificial*, *op cit*.

3 Subrata Dasgupta, "Computer Design and Description Languages", in: Marshall C. Yovits (Ed.), *Advances in Computers, vol. 21.* New York: Academic Press 1982, pp. 91-155; Dasgupta, *Design Theory and Computer Science*. Cambridge: Cambridge University Press 1991; Dasgupta, *Technology and Creativity*. New York: Oxford University Press; Janet Elias and Subrata Dasgupta, "A Cognitive Model of the Engineering Design Mind", in John S. Gero and Nathalie Bonnardel (Eds.), *Studying Designers '05.* Sidney: Key Centre for Design Computing and Cognition 2005, pp. 101-116.

4 Dasgupta, *Technology and Creativity*, *op cit.*

5 Albert E. Musson and Eric Robinson, *Science and Technology in the Industrial Revolution*. Manchester: University of Manchester Press 1969.

6 Michael Polanyi, *Personal Knowledge*. Chicago: The University of Chicago Press

dures, concepts and heuristics that facilitate the creation, manipulation and modification of artifacts.

We can now establish the concept of epistemic complexity in more precise terms. The process of conceiving and bringing into practical form an artifact (*any* artifact) involves the deployment, on the part of the artificer, of his or her knowledge base. Knowledge is, thus, an *input* to the process of artifact creation. But knowledge is also the *output* of that same act: a design embodies one or more operational principles. And in the case of true invention, when the artifactual form is *original* in some significant sense, the operational principles it encodes constitute genuinely *new* knowledge. Thus what distinguished invention or what the engineer-historian Walter Vincenti[7] called "radical design" from "normal design" (also a term Vincenti used) is characterized by two epistemic features: (I) The fact that genuinely new knowledge is produced, predominantly in the form of operational principles; and (II) The fact that old knowledge is put to use in unexpected or surprising way. What seems to most characterize invention or radical design in the realm of artifacts is the *amount, variety and newness of the knowledge embedded in the artifact*. It is this embedded knowledge that I call the epistemic complexity of an artifact.

3. Complexity in normal design

One might expect that there is a direct correlation between systemic and epistemic complexities. Specifically, if an artifact has many components that interact with one another in nontrivial ways and produce behavior that is surprising or obscure, one might expect that such an artifact also encodes a rich body of knowledge. But let us keep in mind that an artifact is epistemically complex not simply because of the amount of knowledge it embeds but the *kinds* of knowledge and the *ways* in which old knowledge combines in the production of the artifact and the new knowledge it generates.

Consider, as an example, the situation Vincenti called *normal design*.[8] As he stated it, in normal design

> The engineer … knows at the outset how the device in question works, what are its customary features and that, if properly designed along such lines, it has good likelihood of accomplishing the desired task.[9]

1962.

7 Walter G. Vincenti, *What Engineers Know and How They Know It.* Baltimore, MD: The Johns Hopkins University Press 1992.

8 *Ibid.*

9 *Ibid*., p. 7.

In normal design, then the overall composition of the artifact is known *a priori*. Brown and Chandrasekaran called this "routine design", and described, in the context of artificial intelligence application, the design of an air cylinder – a piston and rod arrangement which, by moving backward and forward against a spring within a tube creates a to-and-fro movement of some other connected device: air cylinders have a well defined hierarchical form. Starting with this "generic" form, a specific air cylinder may be designed by filling in the details so as to meet specific parametric requirements.[10] In the language of cognitive science, normal design entails the designer summoning up from his personal knowledge system a well-defined *schema* representing the artifact in some stereotypical form, and then *instantiating* this schema to meet specific requirements.[11]

In normal design very little *significant* new knowledge may be produced; old knowledge is used in more or less the same way as in the past. There is little anticipation of surprise. The systemic complexity of the artifact produced by normal design may be considerable but the epistemic complexity will be quite low.

4. The causal connection between systemic and epistemic complexities: An example

A *direct* causal connection between systemic and epistemic complexities can arise in some acts of design and invention. An example is the development of the computer operating system called *Multics* in the 1960s.

In general, operating systems – software that manages computational resources, supports application software and controls the proper functioning of the computer as it goes about its multifarious tasks – is one of the most systemically complex artifacts in the realm of software. Thus when an operating system is conceived and designed to *be significantly original* its systemic complexity directly causes epistemic complexity.

Multics was designed and built at MIT in collaboration with Bell Laboratories and General Electric in the mid-late 1960s as a time-sharing operating system, for the General Electric GE645 mainframe computer.[12] (Later the GE645 and Multics became Honeywell products.) In its mature state Multics consisted of some 1500 modules for a total of approximately one million lines of machine

10 David C. Brown and Balakrishnan Chandrasekaran, "Knowledge and Control for a Mechanical Design Expert System", in: *Computer*, 19, 7, 1986, pp. 92-100.

11 Michael A. Arbib and Mary B. Hesse, *The Construction of Reality.* Cambridge: Cambridge University Press 1986; Roy C. D'Andrade, *The Development of Cognitive Anthropology*. Cambridge: Cambridge University Press 1995; George Mandler, *Cognitive Psychology: An Essay in Cognitive Science*. Hillsdale, NJ: Lawrence Erlbaum Associates 1985.

12 Elliot I. Organick, *The Multics System: An Examination of Its Structure.* Cambridge (Mass.): The MIT Press 1972.

instructions.[13] Its structure was a direct outcome of its overall objective: to create a general computer utility analogous to electric power and telephone utilities which would run continuously and reliably and provide a comprehensive range of services to a population of users interacting with it through remote terminal access. Multics, thus, was conceived as a *technological system*.[14] The particular capabilities that Multics possessed, in response to this overall objective included: (a) time-sharing facilities; (b) an elaborate information storage system that would protect individual user's programs and data from unauthorized access; (c) a sophisticated programming environment for users, including support for several programming languages, inter-user communication facilities (a forerunner of the email); (d) maintenance and monitoring facilities; (e) features to enhance the management of the system's users; and (f) flexibility that would allow the system to absorb new technologies and changes in user expectations.

Clearly, systemic complexity was built into the requirements that Multics would have to satisfy. And though it was not the first time-sharing system to be built – it was anteceded by another system built in MIT called CTSS (Compatible Time-Sharing System, built between 1960 and 1963) and the Cambridge Multiple Access System developed in Cambridge University, England (completed in 1968)[15] – it was the first experiment in creating a comprehensive computer utility. Multics entailed anything but normal design. It had to be *invented* not just designed.

And because it was invented, the systemic complexity inherent in its requirements gave rise to the epistemic complexity of Multics as an artifact.

In fact, its *phylogeny* (that is, its evolutionary lineage) gives us a good sense of this epistemic complexity. It drew upon (a) CTSS; (b) two alternative schemes, invented elsewhere in the early 1960s for implementing "virtual memory", the illusion of unlimited memory capacity;[16] (c) the technique of "multiprogramming" invented almost contemporaneously, whereby several user programs simultaneously share the computer's memory, and the computer's central processor is passed around amongst them so as to keep the processor always busy;[17] and (d) schemes developed in the early-to-mid 1960s for protecting a user's program and data from unauthorized access by other user programs.

---

13 Fernando J. Corbato, Jerome H. Saltzer and Charles T. Clingen, "Multics – The First Seven Years", in: Peter Freeman (Ed.), *Software System Principles*. Chicago: SRA 1975, pp. 556-577.

14 Thomas P. Hughes, "The Evolution of Large Technological Systems", in: Wiebe E. Bijker, Thomas P. Hughes and Trevor J. Pinch (Eds.), *The Social Construction of Technological Systems*. Cambridge (Mass.): The MIT Press 1987, pp. 51-82.

15 Maurice V. Wilkes, *Time Sharing Computer Systems*. London: Macdonald and Janes/ New York: American Elsevier 1975.

16 Peter J. Denning, "Virtual Memory", in: *Computing Surveys* 2, 3, 1970, pp. 153-190.

17 Jack B. Dennis, "Segmentation and the Design of Multiprogrammed Computer Systems", in: *Journal of the ACM* 12 4, 1965, pp. 589-602.

The point is, the designers of Multics did not just draw upon these earlier inventions; they combined, expanded on, and generalized them and in the process created a significantly original product. Furthermore, the development of the Multics system entailed a major experiment in the use of high-level programming languages to write a very large piece of software.[18] It also entailed the application of a design method in which beginning with an initial crude and incomplete system, one used it and observed its behavior, and based on the observed problems the designers simplified, redesigned and refined the system.[19]

Thus, the Multics project both absorbed much prior knowledge and produced significant new knowledge. The artifact itself *embodied* this new knowledge – in the form of what cognitive scientists would call *procedural knowledge*.[20] The situation was rather similar to that of the Britannia Bridge, a wrought-iron tubular railway bridge that crossed the Menai Straits in Wales, built by Robert Stephenson and his associates in the 1840s: here too, the very design and construction of a bridge faced with certain specific requirements produced valuable knowledge about the behavior and properties of wrought-iron structures.[21] The Multics system affords a marvelous case study of an artifact in which systemic complexity is inherent in the desired functional requirements of the artifact, which in turn engendered a rich phylogeny of old knowledge that entered into the invention/design process and generated new knowledge. It is a case study in how systemic complexity gives rise to epistemic complexity.

5. A case of decreasing systemic complexity but increasing epistemic complexity

As a case study in which an artifact has a decrease in systemic complexity but an attendant increase in epistemic complexity, consider another historical episode from computer science. This example also addresses another question: does the evolution of artifacts inevitably entail the emergence of progressively greater systemic complexity?

In fact, there is a general viewpoint that technological evolution carries with it a growth in systemic complexity; that is, artifacts evolve from the simple to

18 Corbato, "PL/1 as a Tool for System Programming", in: *Datamation* 5, 1969, pp. 68-76.

19 Corbato and Clingen, "A Managerial View of the Multics System Development", in: Peter Wegner (Ed.), *Research Directions in Software Technology*. Cambridge (Mass.): The MIT Press 1979, pp. 139-158.

20 Dasgupta, *Technology and Creativity*, *op cit*., p. 37.

21 Nathan Rosenberg and Walter G. Vincenti, *The Britannia Bridge: The Generation and Diffusion of Technological Knowledge.* Cambridge (Mass.): The MIT Press 1978; Dasgupta, "Testing the Hypothesis Law of Design: The Case of the Britannia Bridge", in: *Research in Engineering Design* 6, 1, 1994, pp. 38-57.

the complex, from the less to the more complex. Here, parallels have been drawn between the natural and the artificial since biological organisms are considered to have evolved in complexity.[22] (There is, however, an alternative view of the relationship between complexity and evolution in the natural world[23]).

My case study pertains to the development of the "reduced instruction set computer" (RISC) between 1980 and 1985.

From a functional perspective, a computer presents a certain "façade" to those who are to be its users. This functional façade is usually called a computer's *architecture*.[24] Very briefly (and in somewhat simplified terms) a computer's architecture describes precisely those features of the computer that must be known for a programmer to write an executable program for that machine; it constitutes the lowest-level view of the computer that a programmer can interact with. Examples of architectural features are the details of the computer's instruction set, the syntax and semantics of the instructions, and the types of data that the computer can process.

In general, a computer's architecture expresses one of the basic characteristics of systemic complexity: its various components are mutually dependent; they interact with one another.[25] More interestingly, by the end of the 1970s, the pattern of evolution of computer architectures evidenced a distinct tendency towards *increased* systemic complexity: if one examined a particular genealogical line of computers made by specific manufacturers, one would find that the sizes of the instruction set, the syntax of the instructions, and range of data types, the various modes of referencing instructions and data in memory, etc., had all increased markedly in any manufacturer-defined "genus" of computers.

In the early 1980s, computer scientists at the IBM Thomas J. Watson Research Center, the University of California, Berkeley, and Stanford University independently initiated a movement to *reverse* this trend toward increasing systemic complexity. There were sound empirical and technological reasons for this movement. And based on these arguments, these designers proposed the idea of the "reduced instruction set computer" or RISC – the idea of designing computers with *simplified* architectures wherein all architectural features were greatly reduced in variety, numbers and mutual interactions. The RISC movement represented the notion that evolution in the artificial sciences can proceed towards *decreased* systemic complexity.

However, while the first RISCs that were designed and built were *systemically* simple (relative to their ancestors or their conventional counterparts), the inven-

22 John T. Bonner, *The Evolution of Complexity by Natural Selection*. Princeton, NJ: Princeton University Press 1988.

23 Daniel W. McShea, "Complexity in Evolution: A Skeptical Assessment", in: *Philosophica* 59, 1, 1997, pp. 79-112.

24 Dasgupta, *Computer Architecture: A Modern Synthesis, Volume 1: Foundations*. New York: John Wiley 1989.

25 Dasgupta, *Computer Architecture*, *op. cit*., pp. 108-109.

tion of the RISC concept and the translation of that concept into actual computers were far from being *epistemically* simple. Much historical knowledge was brought to bear by the original inventors in arriving at the RISC concept. And in transforming concept into reality, significantly new knowledge was generated in the realms of computer systems design.[26] The first RISCs were, thus, systemically simple (compared to their predecessors) but such simplicity was gained at the "cost" of considerable epistemic complexity.

6. Epistemic complexity as a marker of the artificer's creativity

What I've tried to demonstrate is that systemic and epistemic complexities are not necessarily coupled. However, our examples also suggest that epistemic complexity is related to the *originality* of artifacts and therefore to the artificer's *creativity*.

An artifact may be systemically complex, but if it is not original, it will be epistemically simple. The products of normal design may exemplify this situation. For example a civil engineer who designs an elaborate flyover system connecting several busy freeways is very definitely creating a systemically complex artifact – both structurally and functionally. But if that system is a product of normal design, it will not be original; no unusual prior knowledge enters into its design or construction, and no new knowledge is produced by it. Epistemically it will be simple.

On the other hand an artifact that is original *will* be epistemically complex, whether they are systemically complex or not. The Multics and RISC systems mentioned earlier exemplify this situation. As another example consider the Italian engineer-architect Pier Luigi Nervi who, in 1936, designed and built aircraft hangars for the Italian Air Force.[27] There were "several traditional solutions" to build such structures: designing aircraft hangars could be seen as exercises in normal design. But Nervi eschewed the normal path. Instead he created an "organism" which transmitted the loads to the supports and columns at the sides and thus provided a large, uninterrupted volume of space for the aircrafts. The huge, dome-like vault was composed of a curved, intersecting network of ribs: this was old knowledge invented eight hundred years earlier by the master masons who built the Gothic cathedrals – but adapted to a radically different type of structure. The sublimity of medieval houses of worship was transposed to the most plebian of buildings – with arresting aesthetic effect. Here was a structure that was

26 David A. Patterson, "Reduced Instruction Set Computers", in: *Communications of the ACM* 28, 1, 1985, pp. 8-21; Manoli G. H. Katevenis, *Reduced Instruction Set Computers for VLSI*. Cambridge (Mass.): The MIT Press 1985.

27 Pier L. Nervi, *Aesthetics and Technology in Building*. Cambridge (Mass.): Harvard University Press 1966.

epistemically complex because it deployed old knowledge in a wholly surprising context. Epistemic complexity is, then, a marker of the artificer's creativity.

7. Descriptors of epistemic complexity

Notice I use the word "marker" above, not "measure". Can epistemic complexity be *measured* at all? For that matter, can *systemic* complexity be measured? In fact, in the latter realm there is no single set of universally accepted measures, each different domain of systemic complexity is perhaps adapted to its own metrics. John Tyler Bonner in his discussion of the evolution of (systemic) complexity of organisms drew upon such measures as body size, diversity of cell types within an organism, diversity of organisms within a community.[28] In the realm of artifacts similar quantitative criteria have been proposed. A well known example is the number of transistors on a integrated circuit chip; the systemic complexity of a software system has been described by the number of lines of instructions, the number of modules comprising the system, the average size of modules, and so on.[29] The study of the (systemic) complexity of algorithms is an important branch of computer science, wherein complexity is measured by the (average or maximum) number of operations of a certain type that the algorithm performs to solve certain classes of problems.[30]

The situation for epistemic complexity is more problematic: it appears to be far less amenable to quantification than its systemic counterpart. One might claim to measure epistemic complexity by simply counting the number of significant and distinct items of knowledge such as facts, concepts, hypotheses, etc., that entered into the invention of an artifact. But such a count would serve as the crudest of measures, not least because what constitutes an "item" of knowledge can be ambiguous. A single "fact" may itself be of limited use in the design or invention process: its significance may only be in its relationship with other items of knowledge – in other words, it may well be an entire *schema* (mentioned earlier) or what cognitive scientists and artificial intelligence researchers call a *semantic network* (that is, a linked network of knowledge and beliefs that show the relationships between the components) that is the significant "item" of knowledge.[31]

For example, in an earlier study, in discussing the invention of the first "superalloy" for gas turbine blades I was able to identify some twenty three significant

28 Bonner, *Ibid*.

29 L.A. Belady and M.M. Lehman, "Characteristics of Large Systems", in: Peter Wegner (Ed.), *Research Directions in Software Technology*. Cambridge (Mass.): The MIT Press 1979, pp. 106-138.

30 Alfred V. Aho, John E. Hopcroft and Jeffrey D. Ullman, *The Design and Analysis of Computer Algorithms*. Reading, MA: Addison-Wesley 1974.

31 Robert M. Harnish, *Minds, Brains, Computers.* Malden, MA: Blackwell 2002.

items of knowledge that appeared to have participated in the invention process.[32] Most of these constituted "old" knowledge which the metallurgists drew upon; the remaining were generated in the process of invention. But such a count conveys nothing of the intricacy of the interactions of these knowledge items, nor the *manner in which they participated in the act of invention*; nor, for that matter, why and how they were invoked at all. The only adequate descriptor of epistemic complexity would be a *description of the ontogenetic process of an invention itself* or some plausible *representation* of this process.

For instance, I attempted to characterize the epistemic complexity of the Britannia Bridge, designed by Robert Stephenson and William Fairbairn in the 1840s by describing a network of cognitive and physical actions involving reasoning, hypotheses construction, experimentation, and model building which Stephenson and his associates engaged in the design process.[33] This description consisted of an interacting web of previously established goals, facts about various bridge forms, general heuristic rules pertaining to engineering design, general problem solving strategies, as well as new facts, new goals, and new hypotheses produced in the course of the design process. Epistemic complexity is, ultimately, a qualitative characteristic: it is not, in general, measurable.

8. Conclusion

In this paper, I have argued that artifacts are characterized by two kinds of complexity. Of these, "systemic complexity" is not unique to artifacts: natural systems manifest it also. The other type of complexity which I have called "epistemic complexity" is uniquely characteristic of human-made systems – artifacts. It is not unique to technological products; "non-useful" things manifest it also. Paintings, sculptures, novels, poems and plays, symphonies, fugues and *ragas* are all infused with epistemic complexity, especially in the intricate ways their creators summon up the past and integrate it into their works.

Understanding systemic complexity tells us *what* the nature of an artifact is. Understanding epistemic complexity tells us *how* that artifact assumed the form it did. Most significantly, in my view, the epistemic complexity of an artifact, useful or otherwise, provides a trace of the artificer's creativity. In this sense it is a *richer* attribute of artifacts than systemic complexity, for it contributes to a depth of understanding of the artifact which analysis of systemic complexity cannot.

If we understand the *sciences of the artificial* as those disciplines that seek to understand artifacts, both in their completed states and the process by which they

32 Dasgupta, *Technology and Creativity*, *op. cit*., pp. 69-74, 152-156.

33 Dasgupta, "Testing the Hypothesis Law of Design: The Case of the Britannia Bridge", *op. cit.*

come into existence, it seems to me that a theory of epistemic complexity has an important place in such sciences.

The Institute of Cognitive Science
The University of Louisiana at Lafayette
PO Box 43772
Lafayette, Louisiana 70504-3772
USA
subrata@louisiana.edu

María José Arrojo

Communication Sciences as Sciences of the Artificial: The Analysis of the Digital Terrestrial Television[1]

Abstract

One new approach to the communication sciences is to analyze them as sciences of the artificial. This involves the consideration of the internal dynamic aspect of the communication sciences from the viewpoint of sciences of the artificial as well as the study of the applied dimension of the communication sciences from the perspective of the sciences of design. Thereafter, the complexity in the communication sciences is taken into account. In this regard, the focus is on the case of Digital Terrestrial Television. The modes of complexity in the communication sciences are seen in the television programming.

1. Analysis of the communication sciences as sciences of the artificial

The communication sciences have a dual character, since they are social sciences and also sciences of the artificial.[2] The character of social science comes from the study a human need purely intersubjective, as communication itself.[3] In this respect, these sciences deal with a type of intentional human actions that take place in a social environment. These disciplines analyze in this way the origin, development and consequences of human action related to the social function of communication.[4] In this regard, they are sciences that take into account the sociocultural environment where this communication of the human agents is carried out.

1 This research project is supported by the Spanish Ministry of Science and Innovation (FFI2008-05948).

2 On this second kind of sciences, cf. Herbert Simon, *The Sciences of the Artificial*. 3rd ed., Cambridge (Mass.): The MIT Press 1996. On the general characters of "science", cf. Wenceslao J. Gonzalez, "The Philosophical Approach to Science, Technology and Society", in: Wenceslao J. Gonzalez (Ed.), *Science, Technology and Society: A Philosophical Perspective*. A Coruña: Netbiblo 2005, pp. 3-49; especially, pp. 10-11.

3 Cf. Roland Backhouse and Philipe Fontaine (Eds.), *The History of the Social Sciences since 1945*. Cambridge: Cambridge University Press 2010.

4 Cf. Robert Clarence Bishop, *The Philosophy of Social Sciences*. London: Continuum 2007.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_26,

Similarly, the communication sciences deal with the expanding of human possibilities of the intersubjective relation of communication. These constant extensions in the communicative field, which are caused by human designs, are studied by these sciences of the artificial. In this respect, what the communication sciences usually do is the "scientification" of a human activity based on a professional experience and develops the existing possibilities of future new assignments.[5] In this way, through the science of the artificial, these disciplines of the communication have new aims, which appear in the designs, using processes to obtain these goals and, finally, obtaining specific results. In which case, it is possible to say that the communication sciences blend with the sciences of design.[6]

As usual with the sciences of the artificial, the communication sciences are orientated to the resolution of specific problems. In this sense, they are applied sciences, since they are intended to solve specific questions within a practical sphere, like human communication.[7] In this way, the communication sciences deal with communication phenomena from different perspectives: the structural and the dynamic.

1.1 The internal dynamic aspect of the communication sciences as sciences of the artificial

In its internal facet, the dynamic aspect of the communication sciences is based on three successive elements: aims, processes and results. These elements endure a practical knowledge that is expressly orientated to the obtaining goals. In this way, the communication sciences, understood as applied sciences, investigate the human aptitude of solving particular communicative problems. Amongst which are those that the television raises and, more specifically, new developments as the Digital Terrestrial Television.[8]

---

5 Cf. Ilkka Niiniluoto, "The Aim and the Structure of Applied Research", in: *Erkenntnis* 38, 1993, pp. 1-21; especially, pp. 8-9; and cf. Gonzalez, "La televisión interactiva y las Ciencias de lo Artificial", in: Maria Jose Arrojo, *La configuración de la televisión interactiva: De las plataformas digitales a la TDT*. A Coruña: Netbiblo 2008, pp. xi-xvii.

6 Cf. Gonzalez (Ed.), *Las Ciencias de Diseño: Racionalidad limitada, predicción y prescripción*. A Coruña: Netbiblo 2007. See especially Gonzalez, "Configuración de las Ciencias de Diseño como Ciencias de lo Artificial: Papel de la Inteligencia Artificial y de la racionalidad limitada", in: Gonzalez (Ed.), *Las Ciencias de Diseño: Racionalidad limitada, predicción y prescripción*, pp. 41-69.

7 Cf. Niiniluoto, "The Aim and the Structure of Applied Research", pp. 1-21; and Niiniluoto, "Approximation in Applied Science", in: *Poznan Studies in the Philosophy of Sciences and the Humanities* 42, 1995, pp. 127-139.

8 Cf. Arrojo, "Caracterización de las Ciencias de la Comunicación como Ciencias de Diseño: De la racionalidad científica a la racionalidad de los agentes", in: Gonzalez, (Ed.), *Las Ciencias de Diseño: Racionalidad limitada, predicción y prescripción*, pp. 123-145.

In the communication sciences one works with designs that are aimed at certain communicative aims. These aims are deliberately chosen, to extend specific communicative human potentials. It happens that these potentials have expanded in recent times. In fact, in the case of television and the new communicative emergent systems (i.e., the internet and the social networks) a new way of broadcasting is taking place. In this case, the considered phenomenon is the study of the audio-visual communication and the deep changes that it is experiencing from the structural and dynamic point of view.[9]

These recent changes arise from an innovative technological development, as it is the arrival of the DDT, which makes new communicative aims possible. These changes result from patterns established by the agents who have some clear and definite aims, among many possible. They are changes that are made by selected processes, finally reaching results, some of which are not always the best. However, these results may be good enough for the aims they were aims at.[10]

The communication sciences that deal with DTT programming is a field of the sciences of design in constant interaction with technology. The area is an artificial net: the elaborated programs, the interactive designed contents, the used channels, the selected time slots, the broadcasting supports chosen, etc. In addition, the programming is orientated to the solution of a certain type of problems. They change depending on the potential audience and the configuration of the company. The processes used and the obtained results are evaluated in terms of *usability* and *usefulness*.[11]

The study of television programming, in artificial reality, corresponds to the science of design. First of all, the study focuses on the aims.[12] The goals established by the programmer can try to obtain the higher possible audience, the best program quality, certain increased economic results, a user experience expanded in time or managing to set a brand image, are some examples. Secondly, the communication sciences, as sciences of the artificial, must analyze the processes followed. In this regard, media developers use the Technology puts at hand (computers, document managing equipment, editing equipment, storage and broadcasting equipment …), that lead to get these ends. And thirdly, it is necessary to consider the results: in what measure programming finds solutions adapted to specific problems, within the framework of expanding human possibilities.

9 Cf. Gonzalez, "La televisión interactiva y las Ciencias de lo Artificial", pp. xi-xvii.

10 Cf. Simon, "Satisficing", in: Douglas Greenwald (Ed.), *The McGraw-Hill Enciclopedia of Economics*. 2nd ed., New York: McGraw-Hill 1993, p. 883.

11 In the case of television programming these two characteristic should constitute ends differentiated in themselves.

12 Depending on the television about which we are speaking, and about the used support, the seeked end will be different.

1.2 The applied dimension of the communication sciences in sciences of design

Nowadays the applied dimension of the "communication sciences" emphasizes that its field of study cannot be reduced to the perspective of the social sciences. Oftenly, its thematic area moves towards more sophisticated designs. At this applied level the use of the new technologies of the information and the communication is particularly important. In this way, all the mass media – including televisions – use some Technologies that allow the creation and diffusion of increasingly elaborate messages.

These factors, which expand our initial communicative capacities – the innate ones, as the human beings we are –, are clearly of artificial nature. They overcome, therefore, the human needs to extend the operational range towards increasingly complex expectations. This involves an increasingly sophisticated applied side in the communication sciences. This practical orientation has been articulated as sciences of the design, which takes into account aims, processes and results.[13]

In this way, in case of the DTT, there are specific communicative designs elaborated to obtain the aims established by the channel. This is made through some processes that seek to solve the problems that appear with the starting of this new communicative environment. All this is aimed at obtaining economic and audience tangible results.

At the same time, the contents of the communication sciences need the technology, which gives what is needed. Because after a creative transformation of the reality,[14] technology offers a device (telephone, television, Internet, etc.) that use as instrumental support for the transmission of the communicative content and the obtaining of the selected aims.[15] Then they start with the design of a communication phenomenon, which takes place on the basis of some aims, provided of the right Technology.

Together with the technological support, these three successive elements – objective, processes and results – entail a practical knowledge: they are orientated towards the obtaining of a goal. According to Simon, when we are dealing with sciences of the artificial, the knowledge goes together with the action.[16] As applied

13 It is a practice where they control the elements that usually analyze the applied sciences when they are sciences of design, since it starts from certain aims in a design (for example, in a programming), that are seeked through a series of processes (the forms of communication) to reach a few results (levels of hearing, managerial profitability, diffusion, etc.).

14 Cf. Gonzalez, "The Philosophical Approach to Science, Technology and Society", pp. 3-49; especially, pp. 11-12.

15 An operative function is a habitual assignment of the technology, cf. Gonzalez, "Configuración de las Ciencias de Diseño como Ciencias de lo Artificial: Papel de la Inteligencia Artificial y de la racionalidad limitada", in: Gonzalez, (Ed.), *Las Ciencias de Diseño: Racionalidad limitada, predicción y prescripción*, p. 49.

16 Simon, *The Sciences of the Artificial*, p. 165.

sciences of design, the communication sciences help to solve present and future problems, so that they combine prediction with prescribing.[17] To such effect they study which solutions have been adapted among all those that the professionals of the sector have used,[18] and they search "scientification" for the guidelines of action.

Therefore, in the scientific field of applied artificial domain, they analyze the new digital systems, as much as the production of communicative contents as the distribution and diffusion. From the scientific perspective, in the case of the DTT it is necessary to analyze if the models used so far in the programming of the analogical television are applicable in this new digital environment and, in that case, what type of changes have to be included to achieve new aims.

Within this emphatic context of the artificial, it is necessary to analyze the communication sciences as sciences of complexity. This can be supported on Simon's ideas,[19] those that – as it is well known – analyze the sciences of the artificial and their complexity. His approach addresses the problem of complexity in an organized and hierarchical internal coordination, which is due to a design. Certainly his approach has not been designed initially for the communication sciences, but I think their approach can shed some light on here.

In fact, the principles of Simon's complexity characteristics, such as internal hierarchy organization, we observe that they are fulfilled in the case of the communication sciences. In which there is an intentional production according to a design, where there is a relationship between their parts and, even, a certain level of hierarchy, so that it is not a chaotic system. Besides the audio-visual products are artificial (man-made) objects, since they arise from the social human action to obtain specific aims, which come to reply to certain problems raised by the society in which they are born.

2. The complexity in the communication sciences: The case of the digital terrestrial television

In the wide field of research named "sciences of the complexity" it is possible to study the communication sciences. Initially, in this field the interaction of disciplines as the physics, the mathematics, the biology, the economics, the engineering and the computer sciences appears, but new possibilities exist.[20] Several authors

17 Cf. Simon, "Prediction and Prescription in Systems Modeling", in: *Operations Research* 38, 1990, pp. 7-14.

18 It is here a parallelism with the applied economics, when he analyzes the management of administration of business firms, an aspect that is relevant for the communication sciences.

19 Cf. Simon, *The Sciences of the Artificial*, pp. 169-181 and 183-216.

20 Cf. Dominique Chu, Roger Strand and Ragnar Fjelland, "Theories of Complexity. Common Denominators of Complex Systems", in: *Essays and Commentaries* 8, 3,

have written about these sciences of the complexity; but, until now, it seems that nobody has included the communication sciences into this set of disciplines that can be analyzed from this perspective.

The component of complexity in the communication sciences appears in several consecutive levels. Firstly, because these disciplines are characterized by their dual character – social and artificial –,[21] inside the empirical field orientated towards the resolution of specific problems. Second, because they are in constant expansion, due to the constant interaction between the internal and external dynamics. This affects both the type of contents they transmit and the devices used in the process. They are, in fact, sciences of design in constant interaction with technology. In third instance, being applied sciences, the communication sciences face the presence of the complexity in three successive stages. They correspond to aims, processes and results.[22]

Due to the first aspect – its dual character: social and artificial –,[23] problems of two different origins are involved in the communication sciences. Attending to the second aspect there is another source of complexity: the communication sciences are empirical sciences, which have an internal type of discipline (contained) and a discipline of external character (of technological, economic and sociological nature). The third aspect – as applied sciences – makes them to be directed to solve specific questions, which are always changing. So they have to be perfectly organized and must be articulated internally, preferably in a hierarchic way.

2.1 The internal articulation of the complexity for Herbert Simon

From an internal perspective, Herbert Simon attention is generally drawn to the background of the complexity. And therefore has studied the problem of the complexity deeply in an organized and hierarchical internal organization[24]. Simon considers that, on having approached the sciences of the complexity, the possibilities of finding answers are much higher if they have a "nearly decomposable" architecture than if the interconnections are less compartmentalized.[25]

2003, pp. 19-30.

21 Cf. Gonzalez, "La televisión interactiva y las Ciencias de lo Artificial", pp. vii-xvi.

22 Cf. Arrojo, "Caracterización de las Ciencias de la Comunicación como Ciencias de Diseño: De la racionalidad científica a la racionalidad de los agentes", pp. 123-145.

23 Cf. Gonzalez, "La televisión interactiva y las Ciencias de lo Artificial", pp. xi-xvi.

24 Cf. Simon, *The Sciences of the Artificial*, pp. 169-181 and 183-216; Simon, "The Architecture of Complexity", in: *Proceedings of the American Philosophical Society* 106, 1962, pp. 467-482. Another influential text is Simon, "How Complex are Complex Systems?", in: Patrick Suppes and Duncan Asquith, (Eds.), *Proceedings of the 1976 Biennial Meeting of the Philosophy of Science Association*, vol. 2. Ann Arbor, MI: Edwards Brothers 1977, pp. 507-522.

25 Cf. Simon, "Complex Systems: The Interplay of Organizations and Markets in Contemporary Society", in: *Computacional and Mathematical Organization Theory* 7, 2, 2001, pp. 79-85.

This factor of interdependence is seen in the communication sciences. It appears in the analysis of television programming: the malfunction of one of the elements that is part of the communicative offer causes a *domino effect*, that take to bad results in the whole program schedule. The schedules are conceived like "a whole." In this way, the audience results of each one of the elements that take part of them depend, to a great extent, on the audience results that the previous contents have obtained. However, each of these elements – the programs – that are part of the schedule is perfectly identified, so they can be replaced by others. In this case, the only negative consequence for the system would be the decrease in the number of viewers caused by the program, if it did not work as the planned way in the set of the schedule, and the economic expenses which one has incurred.

The study of internal dynamics is not enough in the communication sciences. Because, so that this television programming reaches its potential viewers, it has an absolute dependence on the technological component, as well as depending on other environmental factors (business, laws, etc.). The technological component is an external key component here. Because, if the signal – whatever: digital, analogical, cable, satellite … – does not reach its destination due to a failure of the system, an authentic cataclysm takes place in the communicative action, since the possibility of communication doesn't take place.

If one attends to the third stage of the complexity – the aims, processes and results –, then it seems to be clear that the communication sciences, to solve concrete practical problems of the applied sciences, have to be perfectly organized and well articulated internally. Preferably in a hierarchic way, due to the interdependence of the communicative components. All this concerns especially to the internal complex elements. But there are also external factors, besides the technological ones: the broadcasting systems with audio-visual contents have a strong market component. So, one of its main aims is to obtain economic benefits. This determines in a decisive way the processes for the attainment of the aims pursued.[26]

On the analysis of the complexity, Wenceslao J. Gonzalez adopts a wider perspective than Herbert Simon's,[27] who limits himself to a great extent to an analysis of holistic type (about everything and parts). So, together with the structural complexity, a dynamic complexity can occur, where it is necessary to highlight the teleological elements (about ends and means). This means that the complexity is not limited to everything and its parts, but is related with ends and means. But, in addition, the complexity is also to reasons, functions and operations …

In the communication sciences there is certainly an intentional processing according to a design, where a relationship between its parts take place and, even, a

26 Cf. Simon, "Complex Systems: The Interplay of Organizations and Markets in Contemporary Society", pp. 79-85.

27 Cf. Gonzalez, "Complexity in Economics and Prediction: The Role of Parsimonious Factors", in: Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Friedrich Stadler, Thomas Uebel, and Marcel Weber (Eds.), *Explanation, Prediction, and Confirmation*. Dordrecht: Springer 2011, pp. 319-330.

certain level of hierarchy: it is not, in principle, a chaotic system. It is a question, in addition, of artificial objects: the audiovisual products that are born from the social human action to obtain specific aims. The mass are likewise related to the specific problems raised by the society in which they arise: its progress depends on a developed sociocultural context. As the design is more ambitious – as it happens in the DTT or in mobile television – so the difficulties to obtain it will be higher.

2.2 The modes of the complexity according to Nicholas Rescher

From an overall perspective, one can say that "modes of complexity" are multiple, as Nicholas Rescher has emphasized.[28] His analysis insists mainly on two of them: the *epistemic complexity* modes, which affects the complexity in its formulation, and complexity and ontological modes, which correspond to the complexity itself existing in reality.[29] Within the first ones he establishes three manners: a) descriptive; b) generative; and c) computational.

Going from the area of the knowledge to the area of the real thing itself we have several modes of *ontological complexity*: of composition, of structure and as in terms of function. Indeed, Rescher indicates three main ontological modes, they are: (i) the compositional complexity (that attends to the number and variety of the elements that form the matter that is to analyze); (ii) the structural complexity[30] (which is about how the different subsystems of a complex system are structured – the relations can be reciprocal, perfectly coordinated, or a hierarchic relationship of subordination – that in case of the television system is specially relevant (the relationships are hierarchic, so that the channel exercises a position of power with respect to their contents providers, both internal and external); and (iii) the functional complexity (that can be either operational – one that attends to the variety of modes of action or types of functioning – or nomic – that studies the gradual complication of the laws governing the phenomena to be studied).[31]

Rescher's three types of epistemic complexity modes can be illustrated in the case of DTT. 1) The descriptive complexity raised in the specific case of television programming, has to answer three questions: a) what targets you want to reach in each time slot; b) which products; and c) what technological procedure will be used to achieve this. 2) The generative complexity, contemplated before the audiovisual programming, involves a type of program grid instead of generating a sequence. This schedule type will repeated in time depending on audience and economic results and will experience variations depending on both internal and

28 Cf. Nicholas Rescher, *Complexity: A Philosophical Overview*. New Brunswick: Transaction Publishers 1998, pp. 8-16.

29 Cf. Rescher, *Ibid.*, p. 9.

30 What Rescher meant by "structural complexity" corresponds with that described by Herbert Simon when he talks about architecture "decomposable" or "nearly decomposable", in: Simon, "The Future of Information Systems", in: *Annals of Operations Research* 71, 1997, pp. 3-14.

31 Cf. Rescher, *Complexity: A Philosophical Overview*, p. 9.

external conditions. 3) The computational complexity, dealt with in this case, leads precisely to the variety of components of internal type and to factors of external type that need processing. It is the diversity that goes with the television programming, which increases the complexity in this specific area of the communication sciences.

As of these modes of complexity, that Rescher states in general, it is possible to consider the observable communication between humans. To make communication possible, there must be an interaction between issuing agents and other agents considered as receivers. In addition, the technological component takes place in an important way in the most sophisticated communicative process – intervened by the artificial thing. Therefore, the complexity that Rescher calls "functionally" has a very important role in communication processes.

In fact, the receiving agent – individual, group or entity – has to decide what device is going to use to take part in the communicative process (this refers to the functional complexity). For this reason, the audience targeted by the message – the communicative content – needs at least several aspects: a) the message has to be elaborated in a correct way (i.e., under the rules), b) it is broadcasted with suitable procedures and levels of technical quality, and c) the agent must be able to decode the content. So it's understood that in the case of the communication sciences the complexity of the technological component has to be considered.

2.3 Modalities of complexity in the communication sciences: The case of the television programming

Based on the above – Simon's characterization on the types of complexity and Rescher's analysis of the modes of complexity –, the modalities of complexity in the communication sciences are analyzed here in more detail. It is done taking the specific case of television programming, that focuses on the interest raised by the new television: DTT. For this we can start for the factors that Simon identifies when he speaks about "the future of the information systems."[32]

1. The contents are stored and transmitted through television programming. This is similarly to what happens with Simon's information systems, which store and transmit information.
2. According to Simon, the main components in the information systems are two: human and electronic. In the broadcasting systems of audiovisual contents, they are "the human components" the responsible for selecting the aims and designing the necessary processes, so that obtained results are the suitable ones. These agents use the "electronic and technological components" as tools to channel these processes.
3. Similar to what happens in the information systems described by Simon, there is always an organization to coordinate the different stages that audiovisual contents go through. So, these stages – conception, production,

32 Simon, "The Future of Information Systems", pp. 3-14.

planning, exhibition and their possible marketing – often rely on a single organization (or on different companies with the same business and shareholding interests).

4. As in Simon's information systems, in the audiovisual field the business configuration is designed towards aims: making decisions inside the organization must allow to reach the selected aims. Afterwards, the right mechanisms for the production of the contents are established and how these appear to the public.
5. Simon emphasizes the time factor and, above all, the attention as the rarest good in the knowledge society. In the audiovisual area there are an increasing number of channels and supports through which contents can be transmitted. This makes the communication process even more complicated. That is why, the "organizations" change their strategies, the ways to access their contents, and the ways of distributing and programming them. They may even be forced to reformulate their business models.

However, this comparison between Simon's information systems and the television programming as a system of communication is not enough to clarify the complexity. It is suitable to go back to the schemes developed by Rescher on complexity. If they are used to analyze the case of the study of the television programming, the *epistemic* complexity in its descriptive aspect is raised in this case. Because in programming there are numerous parameters that should be specified for their full description. Because the sequence of items during each day, each week or each month can have many modifications or combinations.

There are numerous factors in relation to the generative complexity. In the case of the communication sciences internal type components are specifically affecting the way of making television programs. But there are also external factors, such as the acceptability of the audience. Both combined factors can make that a program or a whole programming schedule is kept or that changes happen (in short, medium or long term).

Some of these decisive elements in the generative complexity are comparative: they depend on what other television channels program at the same time. Others are ingredients of relational type: the season, the day of the week or the time slot that the program is issued, the contents that go immediately before or after, advertisement frames, etc. Other elements are in direct connection with the receiver environment of the television offer, which involves the effectiveness of the promotional campaigns.

There are also factors that affect to the computational complexity, that is, the number of resources to approach the components of internal type and the factors of external type. In this regard, a decisive element is the number of agents in this environment.

But it is within the structural level where the ontological complexity is increased with respect to communication. In the organizational configuration there

are several elements at stake: besides the technological component of base, legal factors intersect with business elements. All of them go together with the properly communicative elements. This set of aspects can be seen clearly in the case of DTT[33] or crossmedia[34] *and* transmedia[35] contents.

It is on this actual side where we see this communication complexity better. In fact, television programming presents a high degree of complexity, especially in the new digital environment. Because, with the proliferation of television channels, the programmer has to look for the maximum return on quantitative and qualitative aspects. It has to be done on each of the channels that depends on it. Some of these channels can share specific programs. Thus, the choice of some contents or others for each of the channels and the corresponding programming will be determined by a variety of factors: (i) the age group to which this program is heading, (ii) what sex it is aimed at, (iii) the status or social class, and (iv) the kind of program (entertainment, educational, informative). Complexity is then given a functional, operational, and nomic type, but within a well defined space-time.

It happens that the development of a program schedule is also determined by the nomic complexity, those who study the gradual complication of the laws that govern the analyzed phenomena, both those related to the communicative process and those external to them. The arrival of the DTT in Spain and in other countries around us has produced an alteration in the regulative internal guidelines of the communication phenomena. The DTT caused a multiplication stage in the supply of audiovisual channels that was not corresponded by a proportional increase of the economic resources for the production of contents, making these internal rules change in an important way.

Together with the above mentioned, which is a clear example of applied science, it is necessary to add the complexity arising from technology. It is something that accompanies the decisive steps of the communicative process: the laws, business and the communication itself. All this occurs inside a society that increasingly demands the use of audiovisual contents through technologically more complex supports.

The viewer wants to control its leisure time choosing what to watch, on what support and at what time. This does not mean that the whole communication system in itself is at random or disordered. It happens that the scheduling policies

33 Cf. Arrojo, *La evolución de la TDT en España: Factores jurídicos, empresariales y de programación*. A Coruña: Netbiblo forthcoming.

34 Transmedia is the notion of narrative that spreads across multiple platforms of mass media, with the aim that every platform contributes with something integral to a total trama *http://www.scriptmag.com/2010/07/28/transmedia-and-writing-starlight-runner-goes-the-distance/* (access on 2/3/2010).

35 Those audio-visual contents that manage to spread in other formats and in other supports, but these do not have sense by themselves, and it is necessary to experience the set of histories in the different platforms to understand them.

are different and they will have to be adapted to the new environment and to the dynamics of consumption of the viewers, as well as of the users of the internet. Each of these supports needs a guideline separated from programming, from communication, from promotion and from business.

In this way new levels of sophistication and of complexity in our understanding of world events are detected.[36] The same thing happens with the analysis of programming or with the study of the communication as a whole. The sum of factors generates an increasingly rich panorama; but, certainly, also more complex. In the communication environment, in general, and television, in particular, economic technological constraints come into play, – not always the programmer is able to access the contents it wishes to form the schedule – and policy development. All these factors add up points of analysis and variables, causing an increase in the complexity of the system and the work of the professionals.

Faculty of Communication Sciences
University of A Coruña
Campus of Elviña
15071, A Coruña
Spain
maria.jose.arrojo@udc.es

36 Cf. William Bechtel and Robert Richardson, *Discovering Complexity*. Princeton: Princeton University Press 1993; and cf. Rescher, *Ibid*.

Team E
The Philosophy of the Sciences that Received Philosophy of Science Neglected: Historical Perspective

Elisabeth Nemeth

The Philosophy of the "Other Austrian Economics"

Abstract

I propose to reconstruct Neurath's early economic theory as a genuinely theoretical, academic contribution to the epistemological controversies which were going on in the not yet well defined field of social science and economics before World War 1, rather than as an early, preparatory stage of his later ideas on socialism (as a planned economy in kind). Emphasizing the difference between his early theory and his later political activism can help us spell out the philosophical impact of Neurath's highly original theoretical approach to economics and how his conceptual innovations there are related to his later contributions to logical empiricism. Tracing Neurath's thought back to the debates on the subject matter of economics and social science before World War 1, also helps us to reconstruct the issues of these earlier debates that disappeared during the "short" 20th century.

The term "The Other Austrian Economics" was coined by Thomas Uebel and refers to a type of economic thought which arose in Vienna in competition with the famous "Austrian School of Economics". This "other" school developed a deeply heterodox approach to economic issues which, at first glance, had nothing to do with its famous counterpart. At second glance, however, it turns out that both shared certain elements.[1] The main representative of these "other Austrians" is Otto Neurath, but Josef Popper-Lynkeus is also an important figure. Although their writings were more or less forgotten after World War 2, a re-appraisal began with Juan Martinez-Alier's book on ecological economics of 1987.[2] Since then, a number of interesting studies on Neurath's economic theories have been published.[3] Many of these studies re-construct Neurath's economic thought from the perspective of the socialist calculation debate of the 1920s and 1930s. This may

1 See Thomas Uebel, "Introduction: Neurath's Economics in Context", in: Otto Neurath, *Economic Writings. Selections 1904–1945* (ed. by Thomas Uebel and Robert S. Cohen). Dordrecht: Kluwer 2004, pp. 1-108. (The volume as a whole will refered to hereafter as "*ONEW*".)

2 Juan Martinez-Alier, *Ecological Economics. Energy, Environment and Society*. Oxford: Blackwell 1987.

3 See, for example, John O'Neill, *The Market: Ethics, Knowledge and Politics*. London: Routledge, 1998, and Elisabeth Nemeth, Stefan Schmitz, Thomas Uebel (Eds.), *Otto Neurath's Economics in Context*. Dordrecht: Springer 2007 (with further references). The volume as a whole will refered to hereafter as "*ONEIC*".

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_27,

suggest a rather straight-forward continuity between the orientation of Neurath's economic theory before World War 1 and the manner it informed his political engagement afterwards. That an important doctrinal continuity exists is undeniable, of course, but I wish to emphasize that Neurath's economic theory took its shape in the *academic* debates in economics and social science before 1918.

1. Can history help philosophy of (social) science?

I would like to begin with a very rough sketch of James Lennox's view on the relationships between science, philosophy of science and history of science.[4] He argues that historical research can play an essential role in clarifying fundamental questions in the sciences, because

> the foundations of a particular scientific field, and … of science generally, are shaped by its history, and to a much greater degree than many of the practitioners of a science realize. There is more conceptual freedom in the way theories – even richly confirmed theories – may be formulated and revised than is usually realized. Studying the way they actually came to be formulated, and revised historically, can be of considerable value in doing philosophical work.[5]

Lennox takes his examples from the theory of evolution and genetics. But it is true not only of biology that there is "more conceptual freedom in the way theories may be formulated and revised than is usually realized". The same can be said about other disciplines, including economics. "A reasonably mature science", Lennox argues, "is the result of a number of decisions made, at various historical nodes, as to which, among a variety of possible options, ought to be taken."[6] Most of those decisions have been forgotten, though, and it is precisely this lack of historical consciousness that characterizes the state of science Lennox called "reasonably mature". In a "mature" science, most of the practitioners agree on the central concepts and methods of their field and therefore do not see any need for reconstructing the possible options that were passed over during the history of their field. Nevertheless, any scientific field has its puzzles and its unsolved problems. In reconstructing the historical origins and development of those problems, philosophers of science may achieve, Lennox argued, a much better understanding

4 Some of the following considerations have been published in E. Nemeth, "Socially Enlightened Science. Neurath on Social Science and Visual Education", in: Mélika Ouelbani (Ed.), *Thèmes de philosophie analytique*, Université de Tunis, Faculté des Humaines et Sociales 2006, pp. 83-112, and in "'Freeing up One's Point of View': Neurath's Machian Heritage Compared with Schumpeter's", 2007, in: *ONEIC*, pp. 13-36.

5 James Lennox, "History and Philosophy of Science: a Phylogenetic Approach", in: *História, Ciêcias, Saúde – Manguinhos*, vol. VIII(3), 2001, pp. 655-669, at p. 657.

6 *Ibid*., p. 659.

of them. Note that, in Lennox's view, better understanding of the *philosophical* problems in a particular scientific field might be achieved by a *historical* reconstruction that situates the current theory in the space of alternative options that were articulated and discussed before the current theory became the dominant one.

> As one traces back through the history of a current theory, one finds various alternatives. This historical research opens up a space of theoretical possibilities that were earlier rejected, or not considered, but in the light of current problems, may seem interesting and suggestive.[7]

From Lennox's point of view, it is not just any alternative theory that showed up at a certain time in history that deserves the philosopher's attention, but primarily those whose foundational problems were discussed by competing scientists before the current theory was accepted.

> [I]t is often true that at that point, those involved in the scientific debate will be quite self-conscious of problems that a couple of generations later submerged as unquestioned, unanalyzed presuppositions of the field's common set of concepts and methods.[8]

Thus, the historical point which Lennox suggests tracing theories back to is the point where scientists themselves still acted, so to speak, as philosophers: when they consciously discussed their conceptual and methodological assumptions. This is not to say, however, that scientists of former periods were *per se* more philosophically-minded than those of later generations. The important point is, rather, that before the basic assumptions of today's "reasonably mature" science were established, scientists had quite a lot to gain from criticizing competing assumptions and from convincing the scientific community that their approaches were sounder than competing ones.

2. Remarks on Neurath's biography

Lennox's reflections can be used as a backdrop against which some interesting features of Neurath's economic theory become visible. During the first decade of the 20th century the debates on methods and value judgements in social science were still going on and polarized many of the younger generation of social scientists in the German speaking world. Neurath was not the only one who thought that the polarisation between the two camps, the German Historical School and the Austrian School, was less substantial than the rhetoric of the debate suggested. For Neurath, however, it was quite natural to look for some sort of integration of the two approaches. He knew both camps rather well. He studied political economy

7 *Ibid*.

8 *Ibid*., p. 667.

at Berlin, the center of the Historical School. After his PhD in 1906, he returned to Vienna and participated in the seminars of some of the main representatives of the Austrian School of Economics. Around 1910 he began to publish on the theory of social science and a wide range of topics in economics and sociology: on the theory of money, the theory of value and political economy, on prize-regulation, sociology of religion and its economic impact, but also on some philosophical and psychological issues, and even on the history of optics. Both the range of topics and the manner in which he discussed them show that Neurath thought of himself as a young scholar about to become a recognized member of the academic community. In 1917 he took an important step towards this by gaining the "venia legendi" in political economy at the University of Heidelberg. Yet it turned out that his academic career ended there. With Neurath having decided in 1918 to join the Social Democrats and to go into politics and having been actively involved in the Bavarian revolution in Munich in 1919, the University of Heidelberg decided to exclude him from its list of lecturers. Although he tried to do so, Neurath was never able to regain a position in academia.

In my view it is important to see that it was not until 1918 that Neurath got involved in politics. Before that, he kept – quite cautiously – his distance from the politics of his day and made a name for himself in the field of economics and social science when their disciplinary borders were not yet established. The time in which Neurath's economic thought was formed was still one in which social scientists acted, so to speak, as philosophers. They disagreed about the nature of what they were doing and the borders of their field of subjects. They articulated their conceptual and methodological assumptions and tried to demonstrate that their approaches were sounder than the competing ones. How ambitious the young Neurath was can be seen from his interventions in these debates. He took every opportunity to address the fundamental conceptual and methodological issues in economics and social science. At the same time he discussed the foundational questions of modern mathematics and physics with Hans Hahn and Philipp Frank. With lessons learned from Ernst Mach and Pierre Duhem in mind, he set out to develop an entirely new conceptual framework of economic theory which he called "calculation in kind".

3. Neurath's early economic theory (1909–1917)

There were two main concerns lurking in the background of Neurath's ambitious theoretical project. The first concerned the divide between the Historical School and the Austrian School of Economics which Neurath thought was a false alternative. Neurath wanted to develop a conceptual framework which was broad enough to include theoretical elements from both sides. On the one hand, Neurath shared with the Austrian School the subjective theory of value and the demand for clearly

elaborated methodological and conceptual standards in economic theory. On the other hand, Neurath appreciated the Historical School for the rich empirical content of their work, for their interest in the economic development of whole populations, and for including certain cultural elements in economic theories.

The second concern that informed Neurath's early approach was that economists had become much too fascinated during the 19th century with exchange-relations under market conditions and price-formation. Their perspective on economic issues was extremely narrow and suggested that only one truly scientific theory of economics was conceivable, namely the theory of market relations as represented in prices. Economic behavior under non-market-conditions became literally un-thinkable. By contrast, Neurath pleaded for a much broader view which, he argued, had also been the view of the classical economists. Smith and Ricardo, for instance, were fully aware of the fact that the relationship between monetary income and real income was deeply problematic and tried to give a theoretical account of these issues.

What was at stake for Neurath was the project of recovering the broader perspective on economics in which the central question was how people become rich or poor. To Aristotle, Smith, Ricardo and other economists, Neurath argued, the subject-matter of economics was "wealth" in all its dimensions. He suggested defining "wealth" as "the totality of pleasure and displeasure that we find with individuals and groups of individuals." It is important to see how Neurath explained why he believed that the term 'pleasure' was particularly appropriate: "The term 'pleasure' has the advantage that in our use of language it comprehends *complex and primitive* facts at the same time".[9] Neurath required a terminology which does *not* invite us to search for the primitive, basic fact to which all other facts can be reduced. (Note that this anticipated a central motive in Neurath's later contributions to logical empiricism: his conception of protocol sentences was, as he once put it, a protest against the idea of basic elementary or atomic propositions.)

A few years later, Neurath changed the terminology to make his intentions better visible, now speaking of "quality of life" rather than of "wealth": "the quality of life is connected with all types of experiences, with eating, drinking, reading, artistic sensibility, religious contemplation, moral speculation, loving, hating, heroic and cowardly behaviour".[10] Remember, however, that the question Neurath wanted to ask is: how do people become rich and poor? To answer this it would not be sufficient simply to give a rich description of what quality of life consists

9 Otto Neurath, "Nationalökonomie und Wertlehre, eine systematische Untersuchung", in: *Zeitschrift für Volkswirtschaft, Sozialpolitik und Verwaltung* 20, 1911, pp. 52-114, reprinted in Neurath, *Gesammelte ökonomische, soziologische und sozialpolitische Schriften* (I), ed. by R. Haller and U. Höfer, Wien 1998, pp. 470-518, at p. 471.

10 Otto Neurath, "Das Begriffsgebäude der Wirtschaftslehre und seine Grundlagen", in: *Zeitschrift für die gesamte Staatswissenschaft* 73, 1917, pp. 484-520. Trans. "The Conceptual Structure of Economic Theory and its Foundations", in: *ONEW*, pp. 312-341, at p. 313.

in. We must also ask how changes of the quality of life come about. So Neurath suggested to reconstruct "whole orders of life", that is, structures the elements of which are as heterogeneous as those we found in "quality of life". "Orders of life" are ensembles of "actions, measures, customs, habits and the like …" which economists have to compare as to their "economic performance".[11]

In a small 1935 monograph of the Vienna Circle series *Einheitswissenschaft* Neurath re-formulated his early economic views again in a slightly different terminology, but the conception remained the same. (Here we see that the logical empiricist Neurath wanted to place his conception of economic theory within the framework of unified science.) And he put his view in a nutshell: "Economic theory deals with the influence particular institutions and actions bear on the standard of living."[12] And in 1938 he published another comprehensive account of his theory as "The Standard of Living".[13] However, it is important to see that the entire conceptual structure was in place before 1918. Neurath called it "calculation in kind" and stressed the theoretical nature of his approach:

> In itself, [calculation in kind] does not represent any one socio-political or economic standpoint; it is merely a way of looking at things. Economic institutions and whole systems of economic organizations can be investigated by the in-kind calculus and it may be found, for instance, that under some circumstances the free market is more efficient than the planned economy… What is essential is how we formulate the problem to be solved. The focus does not lie on the change of prices, of the interest rate, of wages, but on their influence on the satisfaction of needs. Even economic orders that make no use of these concepts may be examined on their efficiency.[14]

After Neurath went into politics in 1918, things changed. In the socialist calculation debate of the 1920s, Neurath defended the view that a socialist economic order had to abolish the monetary system and replace it by a centrally planned economy in kind. (The huge majority of economists did not agree and even the socialist ones were more than sceptical.) So from 1918 onwards, the in-kind-calculus became a tool for the planned economy in-kind that he envisaged and wanted to establish.

---

11 *Ibid.,* p. 318.

12 Neurath, *Was bedeutet rationale Wirtschaftsbetrachtung?* Vienna: Gerold 1935. Trans. "What is Meant by Rational Economic Theory?", in: Brian McGuinness (Ed.), *Unified Science*. Dordrecht: Kluwer 1987, pp. 67-109, at p. 96.

13 Neurath, "Inventory of the Standard of Living", in: *Zeitschrift für Sozialforschung* 6, 1937, pp. 140-151, reprinted in: *ONEW*, pp. 513-525.

14 Neurath, "Die Wirtschaftsordung der Zukunft und die Wirtschaftswissenschaften", Verlag für Fachliteratur, Berlin-Wien 1917, reprinted in: Neurath, *Durch die Kriegswirtschaft zur Naturalwirtschaft*. München: Callwey 1919. Trans. in: *ONEW,* pp. 241-261, at p. 244.

From the very beginning, Neurath was fully aware of the methodological challenge faced by his theoretical approach. Here is one early formulation, put forward during a meeting of the Social Policy Association in 1909:

> Suppose a civil servant has the choice between two places of residence, A and B. In A, he receives a larger quantity of food and accommodation, in B on the other hand a larger quantity of honour. Is it possible to have a calculus such that it summarises for us food and accommodation as one magnitude, and honour as another? Impossible! We are not able to compute such a complex, containing both pleasure and pain, by first separately establishing the magnitude of pleasure, then the magnitude of pain and finally doing the sum. On the contrary, we can only look at such a complex as a whole. Therefore the conversion into money is of no help in this case. … In the end we have to consider a complex of pleasure and pain as *a whole,* if we want to characterise the entire situation of a person.[15]

Note that for Neurath this also held for a whole population.

> The situation is the same if we want to describe the order of life of a people, or of a temporal period, in order to infer from that its favourable or unfavourable conditions. Again we have to look at the entire situation. Here and at many other points as well, the calculus of value reaches its limits, because the value of a sum of goods is not derivable from the sum of the values of the individual goods.[16]

It is important to see that Neurath criticized not only the way economists use the monetary calculus (for which he became notorious among economists). His main intention was to block any attempt to measure a complex structure by using a single unit of measurement. Therefore he also rejected the idea of pleasure units (for which he criticized utilitarian theories), as well as working time units (which some Marxist economists wanted to apply). His main point was not to criticize the use of money, but to raise a more fundamental methodological issue. Its importance becomes clear when Neurath pleaded for the opposite strategy, for beginning with groups of *un*like elements.

> If one begins with groups of like elements, one is all too easily seduced into thinking of the results that one thereby obtains as the only possible ones, and thus into neglecting the analysis of other cases. If we want to investigate groups of elements systematically, we can start out by assuming that each element consists of parts that are fully different from each other.[17]

Neurath's methodological axiom was: construct the subject matter you are dealing with in economics – "wealth", "quality of life" – as an ensemble of heterogeneous

15 Otto Neurath in the general discussion "Über die Produktivität der Volkswirtschaft", in: *Schriften des Vereins für Sozialpolitik* 132, 1910, pp. 599-602. Trans. "Remarks on the Productivity of Money", in: *ONEW*, pp. 292-296, at p. 293.

16 *Ibid*., pp. 293-294.

17 Neurath, "Nationalökonomie und Wertlehre", *op. cit*., at p. 489.

elements; do not presume that its heterogeneity might on a deeper level be reduced to one single element.

It is of considerable interest that a similar methodological challenge plays still a crucial role in modern development economics. In a detailed paper on the conceptual foundations of development studies Sabina Alkire lays much emphasis on the same point. She characterizes "dimensions of human development" as follows: "They are incommensurable, which means that all of the desirable qualities of one are not present in the other, and there is no single denominator they can be completely reduced to (the list cannot be made shorter)."[18] This is exactly Neurath's point.

4. Traces of Mach in Neurath's economic thought

So the central methodological question is how to compare groups of unlike elements systematically with each other. There are different sources from which Neurath drew his inspiration, but we will focus only on one of them. During World War 1 Neurath wrote in a letter to Ernst Mach:

> I have heard with great interest about the latest developments in relativity theory which can be traced to your conception that gravity as a function depends on the total distribution of mass and remains constant toward certain transformations (for example, rotation). It was this idea in your *Mechanics* which has never left me since my first reading, and has influenced my own intellectual development and by indirect paths even in economics. It was your tendency to derive the meaning of particulars from the whole rather than the meaning of the whole from a summation of the particulars, which has been so important. It is in value theory in particular that these impulses have benefited me through indirect paths.[19]

To be sure, Neurath stressed that Mach's influence worked via "indirect paths". Nevertheless, the passage is instructive, not only because Neurath himself related his holistic approach in economics to Mach. It is, I think, a fair interpretation that Neurath wanted to modernize the holistic conception of economics he had inherited from the Historical School by re-formulating it from a Machian point of view. (This was one instance of the transfer of high-level epistemological reflection from physics to economics.) The passage is of interest also because Neurath referred to the chapter of Mach's *Mechanics* in which a new formulation of the law of inertia was given. In doing so, Neurath referred to an important example of the type of reconsideration and reformulation of the basic principles of physics that revolutionized modern physics in the late 19th century.

18 Sabina Alkire, "Dimensions of Human Development", in: *World Development* 30, 2002, pp. 181-205, at p. 185.

19 An undated letter (probably from 1915) from Neurath to Mach, trans. in: John T. Blackmore, Ryoichi Itagaki and Setsuko Tanaka (Eds.), *Ernst Mach's Vienna 1895-1930*, Dordecht: Kluwer 2001, at p. 106.

Mach himself gave an interesting interpretation of what he had tried to do there. In a comment added to the 1908 edition of his *Mechanics*, he stressed that many physicists had come to share his view "that 'absolute motion' is a senseless concept with no content and no scientific utility." However, the issue, Mach continued, is not only to accept this critical insight but to use it in order to "give the law of inertia an understandable sense." In Mach's opinion, there are two ways of doing this. Although the contrast between the two ways is interesting in itself,[20] we will focus only on the one which Mach, following his own interpretation of what he was doing, took "to give the law of inertia an understandable sense":

> the historical and critical way, which considers anew the facts on which the law of inertia rests and which draws its limits of validity and finally considers a new formulation … we must take account of modifications of expression which have become necessary by extension of our experience.[21]

The parallels are clear. While Mach took a fresh look at the facts upon which the law of inertia rests, Neurath took a fresh look at the facts upon which economic theorizing rests. This fresh look conceived of the subject matter of economics as an ensemble of pleasure and displeasure which is influenced by an ensemble of actions and institutions. Neurath also considered the limits of the validity of the economic laws which have been established until now and perhaps a new formulation of them. He aimed at a conceptual framework in which market-exchange could be considered as being only one particular economic order amongst others. In such a framework economists would be able to investigate the effects markets have on the quality of life of particular populations and compare them systematically with the effects which other economic orders would produce. (Neurath suggested investigating and comparing historical ones like the administrative economy of ancient Egypt, war economies of different periods, but also purely theoretically constructed structures.)

For Mach, it was "expanding experience" which made it necessary to introduce modifications of in the formulation of physical law. Neurath's programmatic paper of 1917 developed a conceptual structure in order to allow economists to consider a much broader range of phenomena than previously. In a little thought experiment he indicated the type of consideration that he had in mind, how the economic performance of a particular "order of life" was to be investigated.

> Consider a person who can enjoy two pieces of ripe fruit in the days to come. In one case, the wind blows down the ripe fruit from the tree with the two fruits; in another case, it blows down the unripe fruit, which has to rot uneaten. Then we can say that the initial condition of the wind direction facing the same group of things was more economical in the first case

20 See Elisabeth Nemeth, "'Freeing up One's Point of View'", *op. cit*., at p. 27.

21 Ernst Mach, *Die Mechanik in ihrer Entwicklung, historisch-kritisch dargestellt*. Leipzig: Brockhaus 1883, 6. Aufl. 1908, S. 257. Trans. *The Science of Mechanics*. Chicago: Open Court 1960, at p. 293.

than in the second. We introduced the direction of the wind, so to speak, as an independent variable, assuming that the direction of the wind does not entail any essential differences for the rest of the initial basis of life. … If we could not introduce independent variables, then there would only be the different pleasurableness of total bases of life, but no economic efficiency of individual determining factors.[22]

Note that for Neurath the individual factors which determine the economic efficiency were not bound to be human actions. Later in the text Neurath gave some examples in which the variables are human actions. Nevertheless, it is significant and important that he treated human and non-human variables on the same level. This last feature is directly related to what Ernst Mach says about the "method of variation".

If we have to investigate a set of multiply interdependent elements there is only one method at our disposal: *the method of variation.* We simply have to observe the change of every element for changes in another: it makes little difference whether these latter changes occur "sponaneously" or are brought about through our "will".[23]

For Mach, the method of variation is "the basic method of experimentation", and therefore an essential part of science. (Variation plays also a central role in Mach's famous chapter about thought experiments.) The method of variation has a long tradition in philosophy of science reaching from John Stuart Mill to today's theories of causation. So I think that the method of variation is one promising candidate for further research into the question how far and in what respects Neurath's methodological and epistemological approach to economics followed Mach as its main model.[24]

5. Why should we pay attention to the philosophy of Neurath's economic thought?

The first and maybe easiest answer is that Neurath was a predecessor of economic approaches that became prominent only during the last decades of the 20th century. Today some of the questions Neurath raised are broadly discussed in Ecological Economics, in Welfare Economics and Development Economics. The most im-

22 Neurath, "The Conceptual Structure of Economic Theory", in: *ONEW*, at p. 317.

23 Ernst Mach, *Erkenntnis und Irrtum* (1905), trans. by T. J. McCormack as *Knowledge and Error*, Dordrecht: Reidel 1976, p. 10.

24 There are further places to look for structural similarities with Neurath's economics, e.g., the Machian "elements", Mach's view on the function of thought experiments, his "historic-critical way of looking at things". See Nemeth, "Scientific Attitude and Picture Language. Otto Neurath on Visualisation in Social Sciences", in: Richard Heinrich, Elisabeth Nemeth, Wolfram Pichler and David Wagner (Eds.), *Image and Imaging in Philosophy, Science and the Arts,* vol. 2, Frankfurt: Ontos 2011, pp. 59-83.

portant name here is Amartya Sen.[25] Even if one looks at recent papers from international organisations, it is striking to what extent they deal with the problems Neurath wanted to address.[26] However, I don't think that this first answer can be fully satisfying. What ecological economists call the "incompatibility of values" and what development economists call the "incommensurability of dimensions of societal progress" is indeed closely related to the methodological problems Neurath raised, but the theoretical models of today are much more sophisticated than Neurath's ever were. The same, of course, can be said about Amartya Sen's functions and capability approach. What would be the point of looking in some detail at an earlier, less developed state of the same (or similar) approach?

The second answer I want to suggest therefore is the following: when we read Neurath's economic writings, we see him actively involved in the theoretical development of a particular scientific field. We see him as practitioner of economic science and social science, struggling with some basic notions of his own field and trying to re-conceptualize them. Some of the ideas which we know as Neurath's contributions to logical empiricism are already present in his early economic writings: most prominently the simile of Neurath's boat representing a holistic fallibilism, but also the proto-pragmatic concept of auxiliary motives, his sharp critique of pseudorationalism, his criticism of the fetish of precision etc. In this connection I would like to plead for a sort of "Gestalt-switch". We are used to think of Neurath's early conceptions as markers on his way to logical empiricism, i.e. to what we think of his mature philosophy of science. I suggest that we look at them the other way round: as generalized epistemological concepts which were originally developed and designed with the intention to provide an epistemological basis for Neurath's approach to economics. Relatedly, the way in which Neurath imported some of Mach's ideas into economics may serve as an example of what the unity of science project was meant to be.

The third answer I would like to suggest is related to the way in which Jim Lennox conceives of the relationship between history of science and philosophy of science. When we look at Neurath's economic writings we look at a period in which economics had not yet reached the state of a more or less well defined discipline in its own right. (This state was not achieved until the so-called "high years of theory" during the 1920s and 30s.) The topics that were discussed before World War 1 disappeared. Before then, however, we can see epistemology at work within an emerging scientific field. This should be of high interest to philosophers of science anyway.

---

25 For a discussion of how far the similarities between Neurath and Sen go, see Ortrud Lessmann, "A Similar Line of Thought in Neurath and Sen: Interpersonal Comparability", in: *ONEIC,* pp. 115-130.

26 See, e.g., Enrico Giovannini, Jon Hall, Adolfo Morrone, Giulia Ranuzzi, "A Framework to Measure the Progress of Societies", OECD Working Paper 2009; the Human Development Report 2011 from the UN: Sabina Alkire, "Dimensions of Human Development", *op. cit*., and other papers on poverty measurement by the same author.

Yet we should also remember that economics as it has developed since the 1920s is a child of what Eric Hobsbawm called the "short 20th century" which began with the Russian Revolution in 1918 and ended with the fall of the Berlin wall. The short 20th century was politically, culturally and economically shaped by the tension between socialism on the one hand and liberal democracy and capitalism on the other. Neurath's economic thought developed its profile before that opposition took over the whole political and intellectual world, but the tension between defenders of socialism and capitalism was a main point of discussion already before World War 1. Yet in the young Neurath's day it was still possible to think of a conceptual framework in which a plurality of possible economic orders was conceivable and subject of scientific inquiry, moreover, the borders between economics and sociology were not yet established. Thus – and this would be my fourth answer – tracing Neurath's economic thought back to the debates on the subject matter of economics and social Science before World War 1, allows us to reconstruct this broader range of possible questions and problems that disappeared during the short 20th century. This reconstruction will enrich not only our scientific and intellectual options but also our political ones.

Institute of Philosophy
University of Vienna
Universitätsstraße 7
A-1010, Vienna
Austria
elisabeth.nemeth@univie.ac.at

Veronika Hofer

Philosophy of Biology in Early Logical Empiricism

Abstract

The received view of the influence of Logical Empiricism on the development of the philosophy of biology since 1960s has it that the contributions of the early philosophy of science to this new branch of philosophy are marginal if not detrimental. This perspective is due to misconceptions of some historical contexts and preconditions concerning the beginning of the philosophy of science as a distinctive discipline in the US. This paper contributes to overcoming this received view by identifying some of the main protagonists and their antagonists and the issues and the background of their debates in order to show that Logical Empiricists were interested in the "real" strands of biology for most of the movement's lifetime.

1. Introduction

If one looks for recent scholarship addressing possible connections between Logical Empiricism and philosophy of biology, there are two papers worth considering, both of which call for future work detailing the intellectual activities and relevance of work done by these early advocates of philosophy of science in the realm of philosophy of biology. One is a 2007 paper by Jason M. Byron[1], a young scholar who made a bibliometric survey in order to critically assess what he considered to be the received view, a view according to which the Logical Empiricists both lacked interest in and were irrelevant for the emergence of the philosophy of biology, with the consequence that one had to await the decline of Logical Positivism in the 1960s in order to launch a fruitful philosophy of biology. The other paper, from 2000, by Joe Cain[2] is a good example of this received view which stands out in its critique and careful examination of Betty Smocovitis' book-length argument[3] that there is a model-like influence of Joseph Henry Woodger and the Unity of Science Movement on key actors of the Evolutionary Synthesis (which since its publication in 1996 figures as the standard literature for this claim due to

1 Jason M. Byron, "Whence Philosophy of Biology?", in : *British Journal of Science* 58, 2007, pp. 409-422.

2 Joe Cain, "Woodger, Positivism, and the Evolutionary Synthesis", in: *Biology and Philosophy* 15, 2000, pp. 535-551.

3 V. Betty Smocovitis, *Unifying Biology: The Evolutionary Synthesis and Evolutionary Biology*, Princeton: Princeton University Press 1996.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_28,

the lack of other sources). To provide such missing detail is one of the aims of the present paper and previous works of mine.[4]

2. Arguments for and against the received view

To provide more than one perspective to the argumentation of the received view, let me start with Joe Cain's analysis of Smocovitis' claims as an introduction to Byrons' argument. Discussing Byron's paper in more detail will help to present my own view.

2.1

Joe Cain in his article argues in line with previous papers by David Hull (1969)[5] and Nils Roll-Hansen (1984) [6], that Logical Empiricism had no influence on the community of biologists at large. An earlier paper of Gereon Wolters[7] had the same general thrust. Cain challenges a series of claims which Betty Smocovitis offered in her 1992 book. He attacked in detail Smocovitis' claim that Joseph Henry Woodger (1894–1981) was a key figure linking Logical Positivists with the actors of the Evolutionary Synthesis. Cain put his finger specifically on the lack of historical evidence for any direct social or intellectual connection between Woodger and the key synthesis actors and he criticizes that she is shy to admit this. "Smocovitis leaves as vague the value of any direct social or intellectual connection between Woodger and the key synthesis actors. She works to establish them yet acts as though these connections are irrelevant to her argument."[8] In his paper, Cain strives to assess what these direct intellectual interactions might have been and concludes that all evidence suggests that the connections were very few, whether directly or indirectly. "I know of no positive evidence to suggest any synthesis actors engaged Woodger personally or engaged any of his publications (or followed developments of the logical positivists or the USM). I welcome evidence to the contrary and also welcome specific detail as to how these researchers 'ech-

4 Veronika Hofer, "Philosophy of Biology around the Vienna Circle: Ludwig von Bertalanffy, Joseph Henry Woodger and Philipp Frank", in: Michael Heidelberger and Friedrich Stadler (Eds.), *History of Philosophy of Science: New Trends and Perspectives*, Dordrecht: Kluwer 2002, pp. 325-33.

5 David Hull, "'What the Philosophy of Biology is Not'", in: *Journal of the History of Biology* 2, 1969, pp. 241-268.

6 Nils Roll-Hansen, "E. S. Russell and J. H. Woodger: The Failure of Two Twentieth-Century Opponents to Mechanistic Biology", in: *Journal of the History of Biology* 17, 1984, pp. 399-428.

7 Gereon Wolters, "Wrongful Life: Logico-Empiricist Philosophy of Biology", in: Maria Carla Galavotti and Alessandro Pagnini (Eds.), *Experience, Reality, and Scientific Explanation*, Dordrecht: Kluwer, 1999, pp.187-208.

8 Cain., *op. cit.*, p 537.

oed' each other. Where, in short, is the parallel discourse? Without such evidence, it is fair to conclude few synthesis actors knew directly of this work (Woodger, the logical positivists, or the USM) at the time, and those who probably did know Woodger (perhaps only Haldane and Huxley), knew him only incidentally and did not find his ideas relevant to their own programs."[9]

It might suffice to say that Cain has a point against Smocovitis' claims, but in addition I would like to warn against a narrow understanding of such things as "common ground" and "parallel epistemologies" in collective intellectual biographies. There is reason to believe that J. B. S. Haldane shared special interests in matters of science and society, which brought them in connection with the Logical Empiricists in Vienna (which I will provide later on).

2.2

But let me introduce the arguments of Jason Byron. In his bibliometric analysis, which collected data from four major philosophy of science journals (*Erkenntnis*, *Philosophy of Science*, *Synthese*, and *British Journal for the Philosophy of Science*) covering 1930–1959, Byron aims at a clearer picture of the claim that forms the received view among contemporary philosophers of biology. According to this consensus, the philosophy of science of the 1930s, 1940s, and 1950s was focused only on physics and general epistemology, neglecting analysis of the 'special sciences', including biology. They were focused instead exclusively and simplistically on vitalist issues, which were already dead herrings, or overly impressed by formalisms of the logicians and mathematicians, who wanted to do likewise for biology. And according to David Hull, the call for a return to the science and its special problems would give the field of philosophy of biology its own domain of problems. Therefore, the subdiscipline of philosophy of biology only could have emerged after the decline of logical positivism in the 1960s and 70s. Byron's analysis evaluates the contributions of philosophy of science from the 1930s until the1950s, which assumes that in this period the philosophers of science established for themselves what questions, methods, concepts, and ideas would be central to their discipline or at least worth considering. His results show, that between 1930 and 1959 each year an average of 9,6% of the total number published were on philosophy of biology. 178 articles, which were published in *Erkenntnis* were on philosophy of biology. Furthermore, with this analysis he demonstrates that 40 % of the total of the articles published on philosophy of biology issues did not deal either with vitalism or formal analysis, but were focused instead on "real" philosophy of biology problems. Consequently, Byron not only states that something is wrong with the received view. He instead comes to the conclusion that philosophers of science were interested in biological issues and made them a central part of the discipline. He presumes that given that the Logical Positivists were long interested in "real" problems of biology for philosophy of science throughout

9 *Ibid*., pp. 538-539.

the 1930s, 40s, and 50s, the call to return to the science taken by philosophers of biology in the 1960s, 70s, and 80s, should be understood not as a reaction against, but as a trend which represent instead a broader return to the philosophy of science of the 1930s, 40s, and 50s. The argument of the received view that it required anti-positivism to bloom in order for the modern philosophy of biology to thrive is flawed, because already earlier philosophers and scientists together set the stage for philosophy of science.

3. Context information on the networks between Logical Empiricists and established researchers in the field of biology

In my paper from 2002[10] I show that Woodger came to Vienna in 1926 through a research grant to do experimental work in the Institute of Experimental Biology, "The Vivarium", but his laboratory experiments could not be carried through and he linked up with the intellectuals of the Vienna Circle. Back in Cambridge, Woodger not only studied Whitehead and Russell's new logic intensively, but he also co-founded, in 1932, what became called the "Bio-Theoretical Club". This group of eminent British scientists included Joseph and Dorothy Needham, C. D. Waddington, J. D. Bernal, Dorothy Wrinch, and (occasionally) J. B. S. Haldane. They not only shared the Vienna Circle's epistemological conviction that, after the deep foundational crisis in physics, the intellectual and scientific challenges should be met through theoretical discussions within an entirely different setting. This would include a more egalitarian mode of collaboration in order to strive for a new transdisciplinary scientific unity. Rather than only reviving a "back to workbench program" of naive empiricism, they sought for new strategies to overcome the classical mechanistic and reductionistic views, exhibiting a preference for model building.

It is fair to say, as well as historically correct, that both Woodger and J. B. S. Haldane, the latter a prominent representative for the increasing importance of pencil and paper work in biology, shared an interest in model building, all the more as these theoretical instruments gained favor among scientists who sought help in clarifying epistemological matters within the unfolding complex landscape of genetics. It is certainly important to notice that Neurath approached Julian Huxley as well as Haldane inviting them to participate in the movement's conferences. Haldane answered promptly and positively, but Huxley declined. Bearing this in mind, it is not surprising to find J. B. S. Haldane as a contributor to the proceedings of the second Congress for the Unity of Science in 1936 in Copenhagen with a paradigmatically clear paper about the use of mathematical models in population genetics titled "Some Principles of Causal Analysis in Genetics". This paper ends with a firm confidence in genetic's autonomous disciplinary development and with

10 *Ibid*., pp 328-329.

a demonstratively friendly nod to those in Logical Empiricism who shared a pragmatic standpoint in the dynamics of a discipline.

> I have perhaps given the impression that genetics is a mere name for a series of unsolved problems. On the contrary, it is a highly developed and exact branch of biology, with its own laws and therefore its own mathematics. A geneticist, like a chemist, can do a lifetime of good work without examining his fundamental assumptions. It is essential that genetics should develop independently if the internal contradictions contained in these assumptions are to be laid bare. The geneticist must and will continue to use such expressions as 'a gene for extra wing veins' even though the substitution of this gene for its normal allelomorph only produces extra veins in presence of certain other genes and a certain environment. Nevertheless such expressions, appropriate as they are to laboratory experiments, are less so to agriculture and grossly inappropriate to eugenics. If genetics is to aid in the improvement of the human race, and in the investigation of the nature of life, its assumptions must be rigorously criticized. In this paper I have attempted to criticize a few of its assumptions.[11]

The contrast to the received view is all the more striking when we take into account that the irrelevance of Logical Empiricists for philosophy of biology might be much less due to the prominence of mathematics and logic in Logical Empiricism but rather due to the way the history of the Evolutionary Synthesis was reconstructed and interpreted by its key actors. Sarkar shows, in a long article,[12] how it came that Haldane's central role in the Evolutionary Synthesis was brushed aside. He shows convincingly that it can well be that the "elision of Haldane's role in that history" is more or less a side effect of William Provine's presentation of the vitriolic Fisher-Wright dispute of the 1930s as the most central issue in the history of the evolutionary theory and the effect of Haldane's refusal to take sides. In Sarkar's reconstruction of the history, it comes to the fore that Provine "emerged as a partisan for Wright" in the late 1970s and recommended, also for future historians, to "look upon Wright as the single most influential evolutionary theorist of this century."[13]

Haldane's paper was published next to Nicolas Rashevsky's paper "Physico-Mathematical Methods in Biological and Social Sciences". [14] This too contradicts the received view, especially when we consider that Rashevsky and Carnap collaborated very well in Chicago after Rashevsky's move from the Westinghouse Research Laboratories in Pittsburgh to the University of Chicago was accomplished by the joint effort of the highly established biologist Ralph S. Lillie, the

11 J. B. S. Haldane, "Some Principles of Causal Analysis in Genetics" in: *Erkenntnis* 6, 1936, pp. 355-356.

12 Sahotra Sarkar, "Haldane and the Emergence of Modern Evolutionary Theory", in: Mohan Matthen and Christopher Stephens (Eds.), *Philosophy of Biology*, Amsterdam: Elsevier 2007, pp. 49-86.

13 *Ibid*., p 72.

14 J. B. S. Haldane, "Some Principles of Causal Analysis in Genetics", in: *Erkenntnis* 6 (1936), pp. 346-356; Nikolai Rashevsky. "Physico-Mathematical Methods in Biological and Social Sciences", in: *Erkenntnis* 6 (1936), pp. 357-366.

"most influential evolutionary theorist" Sewall Wright and the famous neurologist Karl Lashley.[15] It may help in overcoming the received view to note that the young genius Walter Pitts was introduced to Rudolf Carnap by Bertrand Russell, who was on a sabbatical in Chicago in 1938. Carnap not only hired Pitts for some jobs but, what is more, he was the one who introduced Pitts to Rashevsky – to the effect that in 1943 Pitts and Warren McCulloch published their famous paper "A Logical Calculus of the Ideas Immanent in Nervous Activity", after Pitts had found a home in Rashevsky's group in Chicago in 1940.

4. The core protagonists in Vienna and Prague

In this section, I will show that local context does matter in the history of science. Let me start with the truism that the Logical Empiricist movement happened to integrate no key figure in matters of biology permanently, let alone a group of philosopher-biologists who led the community discourse; there was no Julian Huxley, no Ronald Fischer, no J. B. S. Haldane around in Vienna, but physicist-philosophers, logicians, economists, such as Schlick, Neurath, Carnap, and Frank. It is also true that the Austrian discourse in genetics and its debates on evolution were influenced by the strong local disposedness for Neo-Lamarckism.[16] But there were figures, nowadays more en vogue due to change in research attitudes and epistemic trajectories, such as Ludwig von Bertalanffy, who sought a way to establish Systems Theory in the late 1920s. It is worth considering that Bertalanffy was welcomed to discuss his ideas about a new organicist holism within the "Study Group for Scientific Cooperation" from 1929 until 1934, led by Rudolf Carnap, which comprised Herbert Feigl, Edgar Zilsel, Karl Polanyi, Egon Brunswik, Paul Lazarsfeld and Wilhelm Reich. And Bertalanffy did not only profit from the discussions with these people, who later would establish considerable reputations in their respective fields, but he was also invited to the Berlin Group around Hans Reichenbach to discuss his ideas contributing to the new trend of systems-thinking in the epistemology of biology with Wolfgang Köhler and Kurt Lewin in the "Society for Scientific Philosophy" which Reichenbach convened at the Charité. The Study Group in Vienna around Carnap as well as Reichenbach's Society intended – as Carnap put it – to foster a better mutual understanding of the disci-

15 Tara H. Abraham, "(Physio)logical Circuits: The Intellectual Origins of the McCulloch-Pitts Neural Networks", in: *Journal of the History of the Behavioral Sciences*, vol. 38, 1, 2002, pp. 3-25.

16 Veronika Hofer, "Der Beginn der biologischen Systemtheorie im Kontext der Wiener Moderne. Diskurslinien und Wissenschaftsgemeinschaften als intellektueller Hintergrund für Ludwig von Bertlanffy", in: Karl Edlinger, Walter Feigl and Günther Fleck (Eds.), *Systemtheoretische Perspektiven: Der Organismus als Ganzheit in der Sicht von Biologie, Medizin und Psychologie*. Frankfurt/Main: Peter Lang, 2000, pp. 137-158.

plines and to clarify their position in the multidisciplinary context of the sciences. This goal can be considered as a sort of propaedeutic for the pluralistic program of the *Encyclopedia of Unified Science*, which would materialize some years later. The debate among the chief organizers Neurath, Frank, Carnap over the respective contributions on biology in the Encyclopedia reveals interesting alliances, which resulted in the preference of the geneticist Felix Mainx over Bertalanffy.[17] The reasons for this will become evident when I reconstruct the disputes in which Mainx was involved with the Berlin biologist-philosopher Max Hartmann.

4.1. Frank's position on biology

Bertalanffy could hope to attract interest in the Vienna Circle, for especially Philipp Frank had an interest in epistemological issues in biology and invited him to conferences. Frank was interested in the various strands of the mechanism-vitalism debate, which implicated moral agendas as well as epistemic ones and which was in need of clarification ever since Hans Driesch turned from his empirical studies in "*Entwicklungsmechanik*" to become a respected advocate and philosopher of neo-vitalism, a widespread trend in German-Austrian biological thinking at that time.

Why in particular was Frank interested in biology? Because Frank had studied biology! This is a fact which deserves attention for the historiography of Logical Empiricism in more than one way. Frank studied biology at the University of Vienna quite extensively. Right after the begin of his studies in Vienna he took a course of five lectures a week in "General Biology" in 1902, followed by a course on cell tissues in his second semester.[18] With a break of five years concentrating on his studies in physics, after having received his PhD Frank again enrolled for courses in biology. He attended a course with Hans Przibram on "Applications of mathematics to biology" in 1907 and again in 1908 with Przibram a course on "Regeneration". He also attended a course on systematics with Richard von Wettstein, a Neo-Lamarckist, an able fighter for cross-disciplinary research programs to test new methods in phylogeny and the father of probably the most influential director of one of the Kaiser Wilhelm Institutes for biology in Berlin Dahlem, Fritz von Wettstein. Frank gained also an excellent insight in current issues of experimental studies in developmental biology through Hans Przibram, the head of the Vivarium, who had studied these problems at length. Right after Frank's last class in the Vivarium, Przibram tried to pin down his own arguments of the crystal analogy in order to explain the regularities in the morphogenesis of organisms. In Vienna Przibram stood out in promoting the indispensability of mathematics to the study of biological phenomena.

Shortly after he finished his studies in biology. Frank published a paper "Mechanism or Vitalism? An Attempt at a Precise Formulation of the Question"

17 Microfiche NR.35, Wiener Kreis Archiv, Neurath, Inv. Nr. 214-215.

18 Cf. Frank's "Nationale" in the Archive of the University of Vienna.

in 1908.[19] One could reasonably speculate about the influence of Przibram on his main argument, for both sharply dismissed the old mechanistic dogmatism in biology. Przibram sought for new ways to integrate methods in chemistry, physics and mathematics into his research on developmental phenomena in animals. In his paper, Frank restructured the vitalists' argumentation respectfully. He treated it as a legitimate question whether the causal analysis in physics suffices for the causal analysis in biology or whether one must supply the causal analysis in biology with extra-hypothesis and special biological constants. But he was quite clear about the requirement that a meaningful statement in biology is always to be formulated as testable causal laws. He concluded that serious work needed to be done by both experimentalists and theoreticians in order to overcome the mix of pseudo-questions and pseudo-problems prevalent in the current biological discourse. This paper was a careful examination, but in its conclusions quite common at his time.

In his 1932 book *The Law of Causality and its Limits*,[20] Frank took pains to criticize Bertlanaffy's positions and presented his Systems Theory as an attempt to dress up vitalism positivistically. In a nutshell: our worldview has to grant randomness a central position, but this indeterminism does not give way for an extramundane will or an entelechy to fill in the local causalities – a criticism that was also directed against Pasqual Jordan and others. Causality is the complementary perspective to teleology, they describe the same matter of fact. Already since Mach and even more with the new physics we must accept that there is no such thing as organization which could be grasped without connection to the things we could know. There is no need for a philosophy of biology based on characteristics like "organization" or "system". The new physics forces us to think of teleology as just another expression of causality, different densities of interactions being characteristic of systems in physics and biology. But there is a positive role for the teleological approach so much in favor with biologists, because this approach considers groups of phenomena which cannot be treated yet with the causal methods of physics. They seek for regularities where they could find them and this is not just legitimate, but always the first step of scientific knowledge. The mental and epistemological strategies to find the crucial evidence for the autonomy of life is and always has been motivated and decided on ideological or political or religious grounds. A counterstrategy would be to break up the isolation of the disciplines to the effect that the ignorance in epistemological foundational problems would make place for better orientation.

4.2. Felix Mainx

Let me briefly introduce Felix Mainx. He was born in Prague in 1900, his father was an officer in the military of the Habsburg Monarchy. After his service

19 Philipp Frank, "Mechanismus oder Vitalismus? Versuch einer präzisen Formulierung der Fragestellung," in: *Annalen der Naturphilosophie* 6 (1908), pp. 393-409.

20 Philipp Frank, *Das Kausalgesetz und seine Grenzen.* Wien: Springer 1932.

in World War I, Mainx studied at the Botanical Institute in Prague under Ernst Georg Pringsheim and Victor Czurda who concentrated on algae, moss, fungi, and other protozoe as objects for studies of the process of speciation. It is important to notice that the kind of studies undertaken by the group around Pringsheim are now considered as important and unifying contributions to the formation of the New Synthesis that took place from 1920s to 1950s. The nature of species and the process of speciation dominated most of the evolutionary studies at this time. They focused on variation, divergence, isolation and selection. Mainx was interested in the evolution of sex determination in algae and protozoa. He was the one in the Pringsheim group to integrate methods used in genetics with methods in physiology. Mainx's work was guided by a new feeling of excitement. This was grounded in a methodological and epistemic change that was accompanied by increasing confidence to explain common problems by new methods in experimental taxonomy, where counting chromosomes, revealing patterns of phenotypic expressions and crossing geographical variances were considered to be a crucial and welcomed import from methods in Mendelian genetics. These studies, performed most noticeable by Dobzhansky and his group in the US, formed one of the major progressive strands in the formulation of the New Synthesis in Evolution Theory.

Mainx won a scholarship to study at one of the three Kaiser-Wilhelm-Institutes in 1928 and 1929. Back in Prague he habilitated in Plant Anatomy and Physiology in 1929, and a second time for Genetics in 1932. Plans for a Department for Genetics at the Institute of Hygiene at the Medical Faculty in Prague failed due to the national socialist occupation of Czechoslovakia and due to his own engagement as a founding member of the "International Society for the Scientific Examination of the Race-Question", an organization agitating against the national-socialist race doctrine. As a consequence, Mainx was forced to resign from his position. He had an invitation for a salaried position with C. D. Darlington at the John Innes Horticultural Research Institution near London, where Pringsheim already had found shelter from racial prosecution in Prague. He received the permit in August 1939, but the outbreak of the war put an end to his emigration plans. Instead he studied medicine and served as a military doctor from 1942 until the end of the war, taking part in several acts of sabotage against the NS-regime that were officially documented after the war. With his father in law, Carl Cori, and his wife he emigrated to Vienna in 1946, and in 1950 he was asked to establish the Institute of General Biology at the Medical Faculty of the University of Vienna, where he taught genetics until 1974.

Frank's preference for Mainx was based on his knowledge of the broader debates as well as first-hand insights into Mainx' experimental work. For the research and the methodological articulations which Felix Mainx, his colleague at the university in Prague performed in the late 1920s and 1930s were part and parcel of the ongoing, increasingly harsh debates between reductionsm in genetics and anti-reductionism in embryology. This discourse provided not only the background of

Bertalanffy's and Woodger's holism, but also for Mainx's studies in evolutionary genetics, which followed of a more pluralistic and integrative research strategy.

As mentioned before, Mainx was invited by Neurath and Frank to contribute to the *Encyclopedia*. The correspondence between Neurath, Frank and Carnap reveals that Frank was firm in preferring Mainx as author of the booklet over Bertalanffy, who was favored by Carnap, with Woodger in the background. Mainx accepted the official invitation to contribute an overview on biology under the title "Foundations of Biology", but it was published only in 1955 due to the troublesome circumstances already outlined and the many post-war delays of the *Encyclopedia* project.

5. Felix Mainx contra Max Hartmann

Max Hartmann was a biologist of highest reputation and director of the three Kaiser-Wilhelm-Institutes for Biology in Berlin. The basic premise of Hartmann's theory of general bi-sexuality was that sex is a universal biological phenomenon and there are always two and only two sexes, which are qualitatively diverse. Not only the Prague group around Pringsheim, but quite many other proto-zoologists did not consider sex as a general and basic phenomenon. For them sexual differentiation was not the primary cause for copulating. They instead held that gamete copulation could also be prompted by their different level of maturity.

These divergent research programs led to a controversy between Mainx and Hartmann that culminated in 1928–1933. Mainx's book *On Sexuality as a Problem of Genetics*[21] contained not only an overview over the state of the art concerning the respective theories and methods, but also a fundamental critique of the epistemological foundation of Hartmann's theory of relative sexuality. With Mainx's book the conflict between the research group in Prague and the research group of Hartmann started as an official affair. As a result of that conflict, Hartmann in 1943 published a book *Sexuality*.[22] The conflict was in a nutshell, that Hartmann's assistant Franz Moewus claimed to have provided experimental proofs for exactly those phenomena, which Hartmann's principle of general bipolar sexuality would explain. The Prague group wanted to repeat his experiments with the same objects, which Moewus for three years had promised to hand over, but this never happened. Not only the Prague group, but all others who tried were unable to corroborate Moewus' findings. So, his interpretation of a chemical connection between gamones and termones, the two kinds of sex substances figuring as the explanans of Hartmann's theory, was attacked on the ground of doubts in his ex-

21 Felix Mainz, *Die Sexualität als Problem der Genetik. Versuch eines kritischen Vergleiches der wichtigsten Theorien*. Jena: Fischer 1933.

22 Max Hartmann, *Die Sexualität: Das Wesen und die Grundgesetzlichkeiten des Geschlechts und der Geschlechtsbestimmung im Tier- und Pflanzenreich.* Jena: Fischer 1943.

perimental credibility. Moewus replied with reference to unsuitable techniques of the opponents and with a heavy anti-Semitic tone against Pringsheim. Hartmann's physiological theory of sex determination, based on the theory of relative sexuality and its biochemical explanation, was clearly opposed to an irritation theory of fertilization, which the Prague group preferred. Moreover, his principle of bisexuality was justified by references to neo-Kantianism.

The debate also attracted the attention of Frank and Carnap in Prague as well as of Hans Reichenbach in Berlin. They invited Hartmann to publish his epistemological point of view, which he did with his "The Methodological Foundations of Biology" in *Erkenntnis* 1932/33.[23] At the end of his paper; Hartmann made it clear that Kantian causality was an a apriori precondition of scientific inquiry both in physics and in biology, and its proper understanding justifies the autonomy of biology both against premature reductions to physical quantitative and the argument that a failure of such a reduction justifies the introduction of vital factors that stand outside the Kantian conception of causality.

Since Mach's influence still loomed large in Prague, it is no wonder that the local botanists gathered around their spokesman Mainx. Mainx's epistemological arguments were in a nutshell: the only criterion of sexuality, based on the epistemological requirements of the economy of thought, is that it is a process from a haploid to a diploid phase and the following reduction. Gametes, sexual organs and individuals with different sexuality should be treated as secondary, because difference is simply the evolutionary result of selection. Interestingly, Mainx's critique also concerned the general induction by means of which the neo-Kantian Hartmann introduced his principles. Mainx held that in this field, in which physiological and genetic methods had to be combined, a clarification of the epistemology was pressing both on a theoretical and an experimental level. He admitted that it is legitimate, as Hartmann did, to aspire to theoretical unification, but stressed that doing so in an almost perfect absence of experimental data, bears dangers for several reasons. The main problem was that Hartmann had introduced bipolarity as an a priori principle from which he deduced the behavior of organisms even in cases where there was a clear indication that the conditions of applying a concept such as sex did not yet exist. In Hartmann's deductive approach the earlier auxiliary concepts, among them sexual potencies, were reinterpreted as a definitive phenotypic effect of a genotypic trait. Thus the physiological concepts were superordinated to the genetic concepts such that sexual properties in the end appeared as something that was given outside the hereditary mechanisms that only occasionally was influenced by genes.

23 Max Hartmann, "Die methodologischen Grundlagen der Biologie," in: *Erkenntnis* 3 (1932/33), pp. 235-261.

6. Conclusion

Let me conclude with the following observations:

First, from the standpoint of the historian of science and of biology, I wish to point to the fact, that it is true that the Logical Empiricists could not integrate biologists of already high international reputation into their local gatherings in Vienna. In this respect the Berlin group fared better due to the participation of the eminent theorists of Gestaltpsychology Wolfgang Koehler and Kurt Lewin. But to claim that out of lack of interest in the current debates of biology they failed to contribute to, and even hampered the field of biology to develop does not do justice to the facts. This is certainly untrue if we consider in particular the engagement in discussing Ludwig von Bertalanffy's developing Systems Approach for a new Philosophy of Biology of the Berlin discussion group led by Reichenbach and the Vienna "Studiengruppe" led by Carnap. As in the field of physics, they understood their enterprise as an urgent philosophical and moral intervention against misconceptions of the consequences of modern science. In the German-speaking countries in the first half of the 20th century biology was loaded with political and moral agendas both in its academic version as well as in its popular versions. It seems that there is a different history waiting to be told, if we consider that the recent historical reconstructions of the 20th century history of biology reveal that the self-celebratory agenda of the architects of the modern theory of evolution overshadowed many other important developments in emerging fields, among them biomathematics and neurology, to which we actually can reconstruct positive connections from Logical Empiricists.

Second, the example of the Mainx-Hartmann debate showed that in their call for a critical examination of the assumptions underlying the hypothesis and leading theories of the KWI for Biology in Berlin, the Prague group engaged, an eminent research laboratory at a historically crucial point. This meant an involvement in "real" biology research practice on many levels, concerned with "real" problems of theory-building in biology. It seems fair then to conclude that the philosophers of Logical Empiricism at least stuck to their program of organizing a different form of doing science together with philosophy.

Third, Frank and Carnap should be recognized for having organized debates concerning biology in the Vienna Circle. I think that with Carnap's "Study Group for Scientific Cooperation" in Vienna, which was organized within the framework of the Verein Ernst Mach, he created a local and novel platform for young scientists. The practice of doing modern philosophy by way of an open collective enterprise intrigued young scientists, as we know from reminiscences from Bertalanffy and others.[24] With Neurath, Frank and Carnap organized the "Second Conference for the Unity of Science" in Copenhagen 1936,which focused on causality and biology. And in the sixth volume of *Erkenntnis* J. B. S. Haldane, and Nicolas

24 Mark Davidson, *Uncommon Sense. The Life and Thought of Ludwig von Bertalanffy (1901–1972), Father of General Systems Theory*. Los Angeles: J. P. Tarcher 1983, p. 50.

Rashevsky contributed important papers on the philosophical ramifications of their work.

Fourth, I wanted to show that it was not just Bertalanffy and a formal logical analysis Woodger-style that Logical Empiricists thought relevant. Instead they were active within the general discourse in the field and, as they had done in physics, not only cooperated with the most prestigious physicists, but reached out to many fellow scientists in their respective local environment. Thus, I approached the revision of the received view – in line with Jason Byron – as a bottom-up analysis of the intellectual interactions between the Logical Empiricists and a few significant biologists in their environment who were willing to discuss epistemological problems of their current research. It is true, these were not the key actors of the Evolutionary Synthesis that figured center-stage in Smocovitis argument.

Let me conclude with the remark that only careful historical research will reveal the particular impact of the cultural environment in which problems are selected as relevant or to be dropped. A case in point are the studies in the history of the Unity of Science Movement published in the last two decades. My own take is, and Jason Byron's as well, that the Logical Empiricists welcomed a variety of research fields, be it genetics, evolution, developmental biology or biomathematics and the early approaches to Systems theory. I also want to underline that concerning the debates of the Logical Empiricists on biology we have to consider the context-relevant selection of problems like vitalism/ mechanism, or concepts like entelechy and inner milieu. This historical selection will only reveal its proper meaning if we take serious Logical Empiricists' pre-World War II aspirations to participate as active as they could in "the reformation of the totality of culture, politics, and reason",[25] ranging from daily life to general philosophical problems.

In line with Byron, I wanted to show that it was not just logical analyses like those of Woodger that they thought relevant. Instead they were active within the general discourse in the field and, as they had done in physics, they reached out to accomplish their goal in cooperation with scientists. Let me also refer to George Reisch's recent findings[26] and agree that a different story can be told once we take their political-historical context seriously. To this kind of work I wanted to contribute.[27]

Department of History of Medicine
Medical University Vienna
Währingerstrasse 25
A-1090, Vienna
Austria
veronika.hofer@meduniwien.ac.at

25 Peter Galison, "The Cultural Meaning of Aufbau", in: (Ed. Friedrich Stadler), Vienna Circle Institute Yearbook, Vol I, 1993, p. 91.

26 George Reisch, *How the Cold War Transformed Philosophy of Science. To the Icy Slopes of Logic.* Cambridge: Cambridge University Press 2005.

27 I wish to thank Elliott Sober for his interest in this research and Michael Stöltzner for helpful discussions.

Julie Zahle

Participant Observation and Objectivity in Anthropology

Abstract

In this paper, I examine the early history of discussions of participant observation and objectivity in anthropology. The discussions resolve around the question of whether participant observation is a reliable method for obtaining data that may serve as the basis for true accounts of native ways of life. I show how Malinowski in 1922 introduced participant observation as a straightforwardly reliable method and then discuss how – and why – most of the discussants in the 1940s and 1950s maintained that the method is reliable only if the researcher takes a whole number of precautionary measures.

1. Introduction

As a distinct research technique, participant observation came into existence around the beginning of the 20th century. Within anthropology, its introduction as method is first and foremost associated with Bronislaw Malinowski. Between 1914 and 1918, he carried out participant observation on the Trobriand Islands, an archipelago east of New Guinea. Based on his findings, he published, in 1922, *Argonauts of the Western Pacific* which provides an account of native life at the islands.[1] In the introduction to the book, Malinowski famously described – and commended – the use of participant observation. In large part, due to his example and promotion of it, the method began to gain ground among anthropologists. It became the defining method of anthropology.

In his presentation of participant observation, Malinowski asserted that the application of the method allows the anthropologist to arrive at an objective account of native life. In the 1940s and early 1950s, anthropologists and other social scientists discussed this and other claims about objectivity. Common to these discussions of objectivity is that they revolve around the question of whether participant observation is a reliable method for obtaining data that may serve as basis for true accounts of native ways of life. The participants in the debate in the 1940s and early 1950s arrived at a slightly different result from Malinowski. Whereas

1 Bronislaw Malinowski, *Argonauts of the Western Pacific*. London: Routledge & Kegan Paul Ltd. 1922.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_29,

Malinowski regarded participant observation as a rather *straightforwardly* reliable method, most of the participants in the later discussion concluded that the method is reliable only if the researcher takes *a whole number of precautionary measures*.

The aim of the present paper is to examine the early history of discussions of participant observation and objectivity in anthropology. I begin by providing an outline of the method of participant observation and different notions of objectivity. On this basis, I first present Malinowski's reflections on participant observation as a rather straightforwardly reliable method. Then I turn to the debate in the 1940s and early 1950s and consider the various reasons advanced as to why the anthropologist must take a whole number of precautionary measures to ensure that the method reliably generates data that may serve as basis for true accounts of native life.

2. The method of participant observation

Malinowski and most of the participants in the debate in the 1940s and early 1950s shared the same conception of the method of participant observation: over an extended period of time, the researcher should participate in the ways of life under study while trying to intervene as little as possible. At the same time, the researcher should observe what goes on around him. Various aspects of this characterization of the method deserve further comment.

To start with the participatory component of the method, the requirement to participate "over an extended period of time" may be rephrased as the demand to carry out participation for approximately two years. This standard was set by Malinowski. Over a period of four years, he spent nearly two and a half years on the Trobriand Islands.

During the long stay in the field, the anthropologist should participate, in the sense of taking part, in the ways of life under study. In the introduction to "Argonauts", Malinowski stressed that he participated in the sense of living among the natives. Also, he pointed out that he sometimes participated in the stronger sense of taking part in the activities of the natives.[2] In general, it is possible to distinguish between various ways and extents to which the anthropologist may participate in the ways of life he studies.

While participating in one way of another, the anthropologist should try to interfere as little as possible in the natives' life. Of course, the anthropologist will inevitably have an impact on the course of daily life when he, say, engages a native in conversation or tries to learn some craft. Also, he may accidentally cause a change of business as usual. Still, and this is the point, he should not actively try to change and interrupt the way the natives normally go about their life. The anthro-

2 *Ibid*., p. 22.

pologist's aim is not to alter the native ways of life that he studies, but to find out about them.

Turning to the observational component of the method, the anthropologist should observe, in the broad sense of taking notice of, what goes on. Above all, this means that the anthropologist should make use of his five senses to register how the natives go about their life. Moreover, "noticing what goes on" is many times taken to include the anthropologist paying attention to, and registering, his own experiences as he is taught, say, how it is appropriate to behave or how to weave a basket.

The whole point of applying the method was famously stated by Malinowski: it allows the anthropologist "to grasp the native's point of view, his relation to life, to realise *his* vision of *his* world".[3]

3. Objectivity

When Malinowski and the participants in the debate in the 1940s and early 1950s reflected on the method of participant observation, the notion of objectivity was often invoked. More specifically, they used the notion in at least three different – though perfectly compatible – senses.

First, objectivity was predicated of the results or accounts based on data gathered by use of participant observation. Here, an objective account of native life was equated with a true account of their life. Accordingly, in the introduction to *Argonauts*, Malinowski talked interchangeably about how participant observation allowed the anthropologist to arrive at the "objective, scientific view of things" and at "the true picture of tribal life".[4] Likewise, this understanding of objectivity informed a passage in *Notes and Queries on Anthropology* from 1951, where it is noticed that by living "outside village territory he [viz. the anthropologist] may be able to take an objective view of the community as a whole".[5]

Second, objectivity was predicated of the method of participant observation. To state that the method is objective was another way of saying that the method reliably produces data that may serve as basis for true accounts of native ways of life. An example of the notion of objectivity used in this sense was provided by Oscar Lewis. He pointed to the concern among some anthropologists with refining participant observation and other methods such that these "might lead to greater precision and objectivity in the gathering, reporting, and interpreting of field data".[6]

3 *Ibid*., p. 25.

4 *Ibid*., pp. 5-6.

5 British Association for the Advancement of Science, *Notes and Queries on Anthropology*, 6th ed. London: Routledge and Kegan Paul Ltd. 1951, p. 41.

6 Oscar Lewis, "Controls and Experiments in Field Work", in: Alfred L. Kroeber (Ed.), *Anthropology Today*. Chicago: The University of Chicago Press 1953, p. 453.

Third, objectivity was predicated of the researcher who carried out participant observation. Thus used, the idea was that the anthropologist who is objective, or who takes an objective stance, is better able correctly to represent what goes on around him when making his observations. For instance, Florence R. Kluckhohn had this sense of objectivity in mind when she related that she had "temporary lapses of cold objectivity" during her fieldwork.[7] And the same goes for Siegfried F. Nadel when he commented that "the observer's personality might easily override the best intentions of objectivity".[8]

When these three notions of objectivity figured in discussions of participant observation, they had one important feature in common: they were invoked as part of examinations of whether participant observation was a reliable method for obtaining data that may serve as basis for true accounts of native life. This formulation was not used in the discussions of participant observation themselves. Still, it captures what is at stake there. In reflections on whether the application of participant observation allows the anthropologist to arrive at an objective picture of native ways of life, what is at issue is whether the method reliably generates data that may serve as basis for true accounts of native life. Similarly, as noticed above, the preoccupation with whether the method is objective amounts to a concern with whether it reliably produces data that may serve as basis for a true picture of native life. Finally, when the objectivity of the anthropologist is in focus, the anthropologist being objective, or mostly so, is regarded as a precondition for the method reliably generating data that may serve as basis for true accounts of native ways of life. Accordingly, the concern with objectivity on the part of Malinowski and the participants in the debate in the 1940s and early 1950s may reasonably be summarized as being at bottom a concern with the question of whether participant observation is a reliable method for obtaining data that may serve as basis for true accounts of native life. This being clarified, it may now be examined what exactly Malinowski and the participants in the debate in the 1940 and early 1950s had to say about this question.

4. Malinowski on participant observation and objectivity

Malinowski's famous presentation of the method of participant observation in *Argonauts* is from 1922. Before, and around, that time, the large majority of anthropologists used other means to gather information about native ways of life. They did not go into the field themselves but had others to collect their data for them. Or, they went into the field yet without living with the natives over extended

7 Florence R. Kluckhohn, "The Participant-Observer Technique in Small Communities", in: *American Journal of Sociology* 46, 3, 1940, p. 343.

8 Siegfried F. Nadel, *The Foundations of Social Anthropology*. London: Cohen and West Ltd. 1951, p. 48.

periods of time.[9] Only a few researchers had carried out participant observation prior to Malinowski and these few researchers had not published anything on their use of the method.[10] Malinowski's introduction to *Argonauts* was the first written piece on participant observations as a scientific method.[11] Against this background, Malinowski "was able to make himself the spokesman of a methodological revolution".[12] Within anthropology at least, the constitution of participant observation as a method came, above all, to be associated with Malinowski.

In the introduction to *Argonauts* Malinowski made it clear that participant observation, as the method was to be called, was conducive to a true picture of native life.[13] In support of this point, he drew attention to several advantages of using participant observation.[14]

Malinowski explained how he participated in the ways of life of the natives in the sense of living among them. As a result, he stressed, he had access to, and was in a position to notice, everything about native life as it unfolded in their village. He made this point in terms of a description of how he would typically pass his day among the natives. Among other things, he commented that "[l]ater on in the day, whatever happened was within easy reach, and there was no possibility of its escaping my notice".[15] Participant observation allowed him to make observations covering all relevant aspects of public life in the native village. Further, Malinowski tells, he insisted on getting access to the more private aspects of native life too. As he stayed with the natives for so long, they ended up accepting this: "as they knew that I would thrust my nose into everything, even where a well-mannered native would not dream of intruding, they finished by regarding me as part and parcel of their life, a necessary evil or nuisance, mitigated by donations of tobacco".[16] In short, Malinowski maintained that he was able to make observations covering *all* aspects of native life.

---

9 Rosalie H. Wax, *Doing Fieldwork. Warnings and Advice*. Chicago: The University of Chicago Press 1971, p. 28ff.

10 Notice that "researchers" should not be taken to include travelers, traders, missionaries, and the like. Here, I shall not address the question of the extent to which some of these may be said to have practiced participant observation before Malinowski did so.

11 Kathleen M. DeWalt and Billie R. DeWalt, *Participant Observation. A Guide for Fieldworkers*. Lanham: Rowman & Littlefield Publishers, Inc. 2002, p. 5.

12 George W. Stocking, Jr., "The Ethnographer's Magic", in: George W. Stocking, Jr. (Ed.). *Observers Observed. Essays on Ethnographic Fieldwork*. Wisconsin: The University of Wisconsin Press 1983, p. 5.

13 K. M. DeWalt and B. R. DeWalt report that, in the sense used here, the term participant observation began to show up in the 1930s. Around 1940, it had gained wide currency – K. M. DeWalt and B. R. Dewalt, *op. cit*., p. 8.

14 Malinowski may also be said to use other means, tied to his manner of presentation, to convince his readers that participant observation is conducive to true accounts of native life. For an analysis of these, see Stocking, *op. cit*., p. 104ff.

15 Malinowski, *op. cit*., p. 8.

16 *Ibid*., p. 8.

Also, Malinowski pointed to another consequence of his long term participation: after some time, the natives got used to his presence and his participation in their ways of life did not have any effect upon their behavior:

> It must be remembered that as the natives saw me constantly every day, they ceased to be interested or alarmed, or made self-conscious by my presence, and I ceased to be a disturbing element in the tribal life which I was to study, altering it by my very approach, as always happens with a new-comer to every savage community.[17]

Thus, Malinowski implied, he was able to observe native life as it really was, that is, as it took place when he was not there.

Lastly, Malinowski related that he participated not only in the sense of living with the natives, but also in the stronger sense of taking part in their activities. Sometimes, he accompanied the natives on their walks, joined them in their games, took part in their discussions, and the like. In this connection, he noticed that "[out] of such plunges into the life of the natives […] I have carried away a distinct feeling that their behavior, their manner of being, in all sorts of tribal transactions, became more transparent and easily understandable than it had been before".[18] In other words, his participation in this stronger sense enabled him to get a better grasp of their ways of life.

Malinowski supplemented these points with a few pieces of general advice: the anthropologist should be thorough and systematic when gathering his data. Further, the anthropologist should remember not to let his personal convictions, views, and the like prevent the data from speaking for themselves: "the main endeavour must be to let facts speak for themselves".[19]

In this fashion, Malinowski presented the method of participant observation as being a rather straightforwardly reliable method for obtaining data that may serve as basis for true accounts of native life. He mentioned only the benefits from using the method. He gives the impression that if the anthropologist keeps his advice in mind, the application of participant observation is plain sailing: its use readily results in data that may serve as basis for true accounts of native life.

5. The debate in the 1940s and early 1950s on participant observation and objectivity

Following Malinowski's promotion of participant observation, it took some time before the method gained wide currency. Likewise, some time passed before papers and chapters, specifically dedicated to reflections on the method, began to be

17 *Ibid*., pp. 7-8.
18 *Ibid*., pp. 21-22.
19 *Ibid*., p. 20.

published.[20] Papers and chapters of this sort mainly started to appear in the 1940s and early 1950s. In these, anthropologists and other social scientists discuss various threats to participant observation as a reliable method for generating data that may serve as basis for true accounts of native ways of life.[21] In particular, they were concerned with six threats. Three of these are versions of the more general problem of missing observations, as I shall call it. The other three threats are versions of, what I shall refer to as, the problem of misleading observations. In the following, I examine the different versions of these two problems in turn. Moreover, as most of the participants in the debate held that the threats may be averted, I look at the proposed solutions to the problems. My primary focus is the methodological reflections advanced by anthropologists. Yet, occasionally, I also refer to papers and chapters by sociologists who participated in the debate and addressed the issues under discussion.

5.1 The problem of missing observations

The general problem of missing observations occurs when the anthropologist's observations fail to cover all the relevant perspectives within, or aspects of, the ways of life under study. If observations representative of relevant perspectives, or aspects, are lacking, this may result in the anthropologist forming a false picture of the native ways of life he studies. The discussion on participant observation in the 1940s and early 1950s particularly focused on three versions of this problem.

5.2 Inaccessible observations

When Malinowski described how he carried out participant observation among the Trobriand Islanders, he conveyed that the natives allowed him to make whatever observations he needed for his study. In the 1940s and early 1950s, anthropologists emphasized that not every researcher is in this fortunate situation. An anthropologist may be denied the possibility of making observations on numerous grounds. These include his sex, his age, his involvement in a conflict among the natives he studies, the way he has explained the purpose of his study, and the natives' sense of privacy. As a consequence, the anthropologist may not be in a position to make various relevant observations.

In the discussion of participant observation in the 1940s and early 1950s, the suggestion about how to respond to this predicament was the following: as far as possible, the anthropologist should monitor his participation so as to be allowed

20 All along, when anthropologists made use of the method of participant observation, they included methodological comments in the introduction to their books on the native ways of life. In the following, these sorts of methodological remarks will not be examined.

21 It is worth noticing that very often these discussions are tied to concrete examples of threats to the reliability of the method. Less commonly, some threat is considered in general or in the abstract.

to make as many relevant observations as possible. In "Notes and Queries on Anthropology", this strategy is exemplified by the advice that "the investigator who hopes to gain a wide view of the culture of a given area must avoid mixing too exclusively with one group".[22] Otherwise, the other groups may not want to talk to the anthropologist. A little later in the same passage, the anthropologist is further encouraged to "pay attention first to that group which is considered 'the best people', and find his informants among them; it will always be easy to work lower in the social scale afterwards; while the reverse may prove impossible".[23]

Needless to say, this strategy, viz. to monitor one's participation with a view to getting access to as many relevant observations as possible, is not always applicable. For instance, there may be nothing the anthropologist can do about his being prevented from making certain observations due to his sex. In these cases, the anthropologist should take his lack of certain types of observations into account when analyzing his data. In that manner, he may try to avoid arriving at a false picture of native ways of life on the basis of his observations.

5.3 Observations not sought out

Another version of the problem of missing observations occurs when the anthropologist fails to seek out accessible situations that put him in a position to make various relevant observations. For instance, Herskovits mentioned how earlier anthropologists wrongly paid attention to the elders only. This meant that they did not seek out situations in which they could have made relevant observations of alternative perspectives within, and aspects of, the ways of life they studied. As Melville Herskovits put it, "for many years it was an axiom of field-work that only the elders could give a 'true' picture of a culture. Today we know better".[24] Other reasons why an anthropologist may commit this sort of mistake are his emotional engagement in the study, his prior field work experience, his personality, and so on.

In the debate in the 1940s and early 1950s, the solution proposed was that the anthropologist should always make sure to seek out all the accessible situations in which there are relevant observations to be made. In this spirit, Herskovits contended that "[t]he best procedure is thus to talk to both men and women, young and old; to observe a wide range of persons in as many situations as possible".[25] This way of dealing with the problem is also exemplified by F. R. Kluckhohn. She wrote: "I wished to evade the bias of viewing the culture entirely from the

22 British Association for the Advancement of Science, *op. cit*., p. 32.

23 *Ibid*.

24 Melville J. Herskovits, *Man and his Works. The Science of Cultural Anthropology*. New York: Alfred A. Knopp 1949, p. 88.

25 *Ibid*.

married-women's perspective. To do this, I had to seek out other acceptable general roles".[26]

5.4 Observations not made

There is also a third version of the problem of missing observation that was mentioned in the debate in the 1940s and early 1950s. It can happen that even though an anthropologist seeks out relevant accessible situations he fails, in these situations, to make various relevant observations. That is, he fails to pick up on, or take notice of, various relevant goings-on within the situations. Again, there may be various reasons for this. One was mentioned by Seymour Miller. He pointed to a situation in which "the observer has become so attuned to the sentiments of the leaders that he is ill-attuned to the less clearly articulated feelings of the rank and file".[27] As a result, the observer does not notice how the rank and file feel.

The proposal about how this problem may be avoided is simple: the anthropologist should make sure to cover all relevant perspectives within, or aspects of, the situation in which he makes his observations. To prepare himself for this, the anthropologist may do various things. For instance, Lewis maintained that the anthropologist should get a firm grip of anthropological theory and method. By acquiring this knowledge, he stated, "we automatically reduce the probability of error".[28] Further, Lewis continued, "to achieve a high degree of objectivity the student must know himself well, be aware of his biases, his value systems, his weaknesses, and his strengths".[29] The underlying idea here is that this puts the anthropologist in a position to prevent his biases, values, etc., from making him overlook relevant perspectives within, or aspects of, a situation. Still, there is always the possibility that the anthropologist does not completely succeed on this account. For this reason, Lewis seemed to suggest, the anthropologist should always include a statement of his interests, assumptions, and the like, in his account of native life. In this way, the reader may know the framework of convictions, assumptions, and the like within which the anthropologist made his observations.

In this fashion, then, the participants in the debate in the 1940s and early 1950s pointed to three versions of the problem of lacking observations. At the same time, they advanced propositions as to how the careful anthropologist may avert these threats to the reliability of the method of participant observation.

5.5 The problem of misleading observations

The other general problem discussed in the 1940s and early 1950s was the problem of misleading observations. It is the following: the anthropologist may make

26 Kluckhohn, *op. cit*., p. 335.

27 Seymour M. Miller, "The Participant Observer and 'Over-Rapport'", in: *American Sociological Review* 17, 1, 1952, p. 98.

28 Lewis, "Controls and Experiments in Field Work", p. 457.

29 *Ibid*.

observations that should not be taken at face value since they are not directly indicative, or reflective, of the ways of life under study. Insofar as the anthropologist wrongly takes observations at face value, he may arrive at a wrong picture of the ways of live he studies. Within the debate on participant observation, especially three versions of this problem were considered.

5.6 The observer's impact on native ways of life

According to Malinowski, his long stay among the natives had the result that his presence ended up having no impact on the natives' behavior whether in public or private. Thus, he claimed, he was able to observe native life as it really was independently of his study. However, in the 1940s and early 1950s, anthropologists maintained that the problem of the observer's impact on the native ways of life is not necessarily solved by the natives' getting more used to the anthropologist. For instance, Benjamin Paul plainly stated that "[t]he presence of the observer influences the event under observation, less so in the case of public and formal performances, more so in the case of informal and private behavior".[30] Consequently, even after some time has passed, the anthropologist cannot assume that his observations are directly indicative of native life as it takes place when he is not there.

The suggested response to this problem was that the anthropologist should try to determine to what extent, and in what ways, his presence has an effect on the natives' behavior. Paul continued by exemplifying this line of approach when he wrote that "[c]ases of domestic quarrels that are witnessed, for instance, should be compared with reports of quarrels that are not witnessed by the investigator".[31] Once the anthropologist has an idea of the extent and nature of his impact on the natives' behavior, he may then take this into consideration when using his observations as basis for an account of the natives' ways of life. In that manner, he may avoid taking the misleading observations at face value.

5.7 Natives' incorrect accounts

When an anthropologist participates in the native ways of life, he will typically have conversations with the natives and he will overhear them talking to each other. In the debate in the 1940s and early 1950s, it is stressed that the anthropologist should not always take the natives' accounts at face value: the natives may, intentionally or nonintentionally, provide incorrect representations of their ways of life. An instance of this problem was reported by Nadel: "[o]ften I have been told by Nupe noblemen that some religious cult of the peasants was merely a ridiculous

30 Benjamin D. Paul, "Interview Techniques and Field Relationships", in: Alfred L. Kroeber (Ed.), *Anthropology Today*. Chicago: The University of Chicago Press 1953, p. 443.

31 *Ibid*.

and nonsensical practice, not worth recording. Where my exalted informants *did* recall it, their description was full of misunderstandings and distortions."[32]

The proposal advanced as to how the anthropologist may avoid wrongly taking the natives' accounts at face value is that he should try to determine whether, or to what extent, the natives' accounts correctly portray their ways of life. There are various manners of doing so. For example, Nadel pointed out that

> [i]nasmuch as these forms of bias are also sources of error, they can be checked and controlled by various means – by the judicious choice of informants from various walks of life; by a judiciously concrete technique of questioning; by the collection of several complementary statements and of numerous case studies; above all by ascertaining the 'bias' which must follow from the general organization by society.[33]

By checking the natives' accounts in this manner, the anthropologist may take their correctness into consideration when using them as basis for a portrayal of the ways of life he studies.

5.8 The observer's distortion of the situation

A third version of the problem of misleading observations, discussed in the 1940s and early 1950s, occurs when the anthropologist distorts what is going on in the situation he observes. There may be various reasons why this happens. For example, Morris S. Schwartz and Charlotte G. Schwartz noticed that if the researcher "is studying the authority and power relations in a social structure, his own difficulties in accepting authority or wielding power may prevent him from seeing the situation realistically".[34] Obviously, insofar as the anthropologist's observations are distorted, they should not be taken at face value.

The suggested response to this problem was that the anthropologist should take steps to ensure that his observations will not be distorted. Among other things, the anthropologist should acquire a broad knowledge of anthropology, avoid developing too close emotional ties with the natives, and examine his values and convictions. On this basis, he may try to reduce the distorting impact of factors like these. In this connection, Schwartz and Schwartz commented that "discovering one's biases becomes a continuous process of active seeking out and grappling with one's limitations and blocks […] the more perspectives from which we see the bias, the greater the possibility of minimizing its effects".[35]

Thus, the participants in the debate in the 1940s and early 1950s considered three versions of the problem of misleading observations. And, like in the case of the problem of lacking observations, they made suggestions as to how the careful

32 Nadel, *The Foundations of Social Anthropology*, p. 38, italics in original.

33 *Ibid*., p. 39.

34 Morris S. Schwartz and Charlotte G. Schwartz, "Problems in Participant Observation", in: *American Journal of Sociology* 60, 4, 1955, p. 351.

35 *Ibid*., p. 353.

anthropologist may prevent the different versions of this problem from undermining the reliability of the method.

6. Conclusion

In the present paper, I have examined the early history of discussions of participant observation and objectivity in anthropology. These discussions revolve around the question of whether participant observation is a reliable method for obtaining data that may serve as basis for true accounts of native ways of life. First, it was shown how Malinowski regarded participant observation as a rather *straightforwardly* reliable method. Next, the debate on the method in the 1940s and early 1950s was considered. It was demonstrated how – and why – most of its participants maintained that only if the anthropologist takes *a whole number of precautionary measures* is participant observation a reliable method for generating data that may serve as basis for a true picture of native life. Of course, the debate on participant observation and objectivity did not end there: it carried on and it is still ongoing. It is notable that in the early discussions reviewed here the ideal of scientific objectivity and its applicability to anthropology was taken at face value, whereas this has been questioned with increasing frequency in more recent times. A survey of the further development of the discussion and an investigation of how this may be related to developments in the philosophy of social science generally, however, is the topic for another paper.

Department of Media, Cognition and Communication
Section of Philosophy
University of Copenhagen
Njalsgade 80
2300, Copenhagen
Denmark
jzahle@hum.ku.dk

Jean-Marc Drouin

Three Philosophical Approaches to Entomology

Abstract

The first philosophical approach to entomology deals with insects as small animals. Due to the physical differences linked with the difference of scale, small animals seem to live in another world. The second approach deals with nomenclature and classification. It shows the progressive making of the concept of insect. In this process, groups like the crustaceans or the spiders were split off from the insects. The third approach deals with the interpretation of insect societies. – Insects can offer to philosophy a set of thought experiments, which have not been fully exploited.

1. Introduction

The phrase "a philosophical approach to entomology" sounds slightly paradoxical. Many philosophers would agree with Buffon saying in the *Discours sur la nature des animaux*, published in 1753, that a fly must not take more space in a naturalist's mind than it takes in Nature.[1] This spiteful remark was aimed at Réaumur, Buffon's main rival. Réaumur was the author of the celebrated *Mémoires pour servir à l'histoire des insectes*, in six volumes (1734–1742).[2] Peculiarly, Buffon despised Réaumur's admiration of the architecture of the bee's cells; he stated, as a principle, that the less one reasons, the more one admires. Buffon considered the shape of the bees' cell as the mechanical result of the forces of a huge number of bees. By taking into account the distance between insects' behaviour and man's behaviour, Buffon avoided one anthropomorphist bias. But he did not avoid the anthropocentric bias of placing the human species at the centre of nature. Anthropocentrism and anthropomorphism look like the Charybdis and Scylla of natural history, but while anthropocentrism is a metaphysical and ethical point of view, anthropomorphism can offer a useful metaphor. For instance when the French entomologist Jean-Henri Fabre described an insect as a craftsman he did it to express

1 Georges-Louis Leclerc Comte de Buffon, *Histoire naturelle*, *générale et particulière,* vol. IV. Paris: Imprimerie Royale 1753, p. 92. (See also a web edition, <www.buffon.cnrs.fr>)

2 René Antoine Ferchault de Réaumur, *Mémoires pour servir à l'histoire des insectes*. Paris: Imprimerie Royale, 6 vol.1734-1742.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_30,

what we nowadays call the ecological function of this insect and to vividly depict the succession of gestures necessary to perform its task.[3]

Despite its importance, the controversial problem of anthropomorphism is not the only one. The current paper will be devoted to three other approaches to insects. These approaches are termed "philosophical" because they represent broad motifs of inquiry that set its scene by determining what type of observation or distinction is deemed worthy of note.

2. The size of insects

The first philosophical approach to entomology deals with insects as small animals. Michelet, for instance, in *L'Insecte* – one of the popular books of natural history, which the French historian wrote in co-operation with his wife Athénaïs Mialaret – depicts beetles wearing their heavy carapace as an armour of the Middle Ages and fancies that it is proportionally like a man bearing the obelisk of Luxor.[4] Popular science often explains that a flea, if it had the same size as a man, could jump on the top of the Eiffel Tower. Many authors have shown, from a mathematical point of view, the fallacy of this kind of comparison. The strength of an animal is as the square of its length, while the weight is proportioned to the cube of its length. But the computation cannot refute the dream and the miniature world of insects still strongly appeals to the imagination.

Due to the physical differences linked with the difference in size, small animals seem to live in another world.

The problem was known by Galileo, but it has been forgotten by metaphysicians.[5] Hence, the famous meditation by Pascal, who, in the *Pensées*, after depicting the earth lost in the universe, focussed on a "ciron" (a cheese mite) and

3 Jean-Henri Fabre, *Souvenirs entomologiques*. Paris: Delagrave, 10 vol. 1925. See for instance the case of the Dung Beetle, VI, 13, p. 243. On Fabre, see: Yves Cambefort. *L'œuvre de Jean-Henri Fabre*. Paris: Delagrave 1999; Patrick Tort, *Fabre. Le miroir aux insectes.* Paris: Vuibert/ADAPT 2002.

4 Jules Michelet, *L'Insecte*. Paris: Hachette 1857, p. 133. See also the new edition by Paule Petitier, Sainte-Marguerite sur Mer: Edition des Equateurs 2011. The other naturalist books by Michelet are: *L'Oiseau* 1856; *La Mer* 1861; *La Montagne* 1868.

5 Galileo Galilei, *Discorsi e dimostrazioni matematiche, intorno a due nuove scienze* [*Discourses and Mathematical Demonstrations Relating to Two New Sciences*]. Leiden: Louis Elsevier 1638. See also John Burdon Sanderson Haldane, "On Being the Right Size", in: John Maynard Smith (Ed.), *On Being the Right Size and other Essays.* Oxford: Oxford University Press, [1927] 1985, pp. 1-8. See also Augustin Cournot *Matérialisme. Vitalisme. Rationalisme. Etudes sur l'emploi des données de la science en philosophie*. Paris: Hachette 1875. Reissue by Claire Salomon-Bayet: Paris, Vrin 1987. See also Jean-Marc Drouin, "Quelle dimension pour le vivant ?", in: Thierry Martin (Ed.), *Le tout et les parties dans les systèmes naturels*. Paris: Vuibert, pp. 107-114.

imagined a universe in the "ciron"…[6] The first chapter of *On Growth and Form,* the masterpiece of D'Arcy Thompson (first published in 1917) is an answer to this kind of meditation.[7] As Stephen Jay Gould says in his preface to a recent reissue: "the author descends from the ordinary gravitational world of our own species through the realm of surface forces inhabited by insects, to the utterly unfamiliar domain of bacterium".[8]

It has often been stated that nothing is small or large *per se*. D'Arcy Thompson analyses and refutes this commonplace:

> We are accustomed to think of magnitude as a purely relative matter … and we are apt accordingly to suppose that size makes no other or more essential difference, and that Lilliput [the country of the little people in Gullivers' travel] and Brobdingnag [the country of the giants] are all alike, according as we look at them through one end of the glass or the other.[9]

3. Nomenclature and classification

The second philosophical approach to entomology deals with nomenclature and classification.

To classify can just be to set things in an order (for instance the alphabetical order) which makes it easy to find them. It can be sorting things according to their real, or supposed, utility or harmfulness. But the entomologists, like all other naturalists, did not content themselves with those practical classifications, they looked for a classification which caught something of the order of nature. The history of entomology shows the progressive making of the concept of insect. In Linnaeus' work, the group of the insects included butterflies, moths, flies, mosquitoes, bugs, bees, ants, wasps, beetles, fleas, grasshoppers, dragonflies, and all kinds of animals which are still considered as insects, but it also included crabs, shrimps, lobsters, crayfishes, woodlice, spiders, scorpions, mites, centipedes, and all kinds of animals, which are no longer considered as Insects.[10] In the second half or the 18th century and the first half of the 19th century, many groups were excluded from the insects. For instance, all the species which do not have six legs. All these efforts of classification resulted in a phylum, the arthropods, divided in several classes: insects, arachnids, crustaceans, centipedes.

6 Blaise Pascal, *Pensées,* in: *Œuvres complètes*, Jacques Chevalier (Ed.), Paris: Gallimard 1954.

7 D'Arcy Wentworth Thompson, *On Growth and Form*. Cambridge: Cambridge University Press 1961, pp. 15-48 [posthumous and abridged edition by John T. Bonner]. Reissued with a preface by Stephen Jay Gould, Cambridge: Canto Editions 1992.

8 Stephen Jay Gould, Preface to Thompson, *On Growth and Form*, *op. cit.,* p. x.

9 D'Arcy Wentworth Thompson, *On Growth and Form*, *op. cit.,* pp. 16-17.

10 Carl von Linné, *Systema naturae,* 10th issue, Stockholm: 1758. See also Mary P. Windsor, "The Development of the Linnean Insect Classification", in: *Taxon* 25, 1, 1976, pp. 57-67.

The entomologists also dealt with the "orders" into which they divided the class of insects: for instance, coleoptera (beetles) or lepidoptera (moths and butterflies). The orders are divided into families, the families into genera, the genera into species. The Linnean nomenclature, still in use nowadays, names each species by its generic name completed by a specific attribute.

In order to construct all these classifications, the entomologists tried to use the natural method of classification, taking into account several characteristics of different organs. This method was initiated by the French botanists. It was introduced in entomology circa 1800, by Pierre-André Latreille.[11]

At first, the Darwinian revolution was chiefly a reinterpretation of the natural classification. In the *Origin of Species* Darwin argues that the descent with modification is the principle which can justify the rules which the naturalists obey in their day to day work. For instance the importance for classification of rudimentary organs is explained by a comparison "with the letters in a word still retained in the spelling, but become useless in the pronunciation, but which serve as a clue in seeking for its derivation".[12]

In the late 20th century a new trend in classification, cladistics, was proposed by the German entomologist Willi Hennig. The principle is to construct a strictly and only genealogical classification.[13]

4. Political debates on insect societies

The third philosophical approach to entomology deals with the interpretation of insect societies.

4.1 Kings or queens?

During many centuries, the main question was whether the ant colonies or the beehives are ruled by kings or by queens. One of the more significant texts from this point of view is *The Feminine Monarchie or the History of Bees* by Charles Butler. First issued in 1609, the book was often republished.[14] It is a genuine guide of beekeeping, which also contains political views. The hive is pictured as an Amazonian kingdom and the drone "is but an idle companion, living by the sweat

11 Pierre-André Latreille, *Considérations générales sur l'ordre naturel concernant les classes des crustacés, des arachnides et des insectes*. Paris: Schoell 1810.

12 Charles Darwin, *On the Origin of Species*. London: John Murray 1859. Reprint, Cambridge (Mass.): Harvard University Press 1964, p. 455.

13 Willi Hennig, "Phylogenetic Systematics", in: *Annual Review of Entomology*, vol. X, 1965, pp. 97-116. French Translation by Daniel Gouget et al., published in: *Biosystema*, 2, 1987, pp. 1-30.

14 Charles Butler, *The Feminine Monarchie or the History of Bees*. London: John Haviland 1623 (first published 1609).

of others brows".[15] Butler warned the reader that things are different in the human species. Anyway the issue of the gender of the leader of insect societies was settled when the Dutch naturalist and physician, Jan Swammerdam, observed with a microscope the genital organs of bees and ants.[16]

4.2 Monarchy or republic?

Not a hint of this entomological knowledge can be found in Bernard Mandeville's famous political writing, *The Fable of the bees or private vices, public benefit*s. Published in 1714, Mandeville's fiction showed a hive whose bees having decided to be virtuous became so poor that they "flew in a hollow tree".[17] Mandeville's hive was just an image of a rich city, whose economical activity is founded on greed and fraud.

The image of insect societies has often been controversial.[18] But during the French Revolution, the controversies gained a great importance because of the political debates about monarchy and republic and because of the economical need to produce wax and honey. In any case, it can be said that the entomologists em-

15 Butler, *op. cit.,* first page of chapter IV. On Butler and the gender issue, see: Frederick R. Prete, "Can Female Rule the Hive? The Controversy over Honey Bee Gender Roles in British Beekeeping Texts of the Sixteenth-Eighteenth Centuries", in: *Journal of the History of Biology*, XXIV (1), 1991, pp. 113-144.

16 Jan Swammerdam, *Histoire naturelle des insectes*. Utrecht: Ribbius 1685 (French translation).

17 Bernard Mandeville, *The Fable of the Bees or Private Vices, Publick Benefits.* With a Commentary Critical, Historical and Explanatory by F.B. Kaye. Oxford: Clarendon Press 1924 (Reprint; Indianapolis: Liberty Classics 1988).

18 Several publications deal with the controversies about Insect Societies. See for instance: Perru, "La problématique des insectes sociaux: ses origines au XVIIIe siècle et l'œuvre de Pierre-André Latreille", in: *Bulletin d'histoire et d'épistémologie des sciences de la vie*, vol. X, 1, 2003, pp. 9-38; Marc Ratcliff, "Naturalisme méthodologique et science des mœurs animales au XVIIIe siècle", in: *Bulletin d'Histoire et d'Epistémologie des sciences de la vie*, vol. III, 1, 1996, pp. 17-29; Jean-Marc Drouin, "L'image des sociétés d'insectes en France à l'époque de la Révolution", in: *Revue de Synthèse*, vol. IV, 1992, pp. 333-345; Jean-Marc Drouin, "Ants and Bees between the French and the Darwinian Revolution", in: *Ludus Vitalis*, vol. XII, 24, 2005, pp. 3-14; Sarah Jansen, "Ameisehügel, Irrenhaus and Bordell: Insektenkunde und Degenerationdiskurs bei August Forel (1848-1931). Entomologe. Psychiater und Sexualreformer", in: Norbert Haas, Rainer Nägele and Hans-Jörg Rheinberger (Eds.), *Kontamination*. Eggingen: Edition Isele 2001, pp. 141-184; Abigail Lustig, "Ants and the nature of nature in August Forel, Erich Wasmann and William Morton Wheeler", in: Lorraine Daston and Fernando Vidal (Eds.), *The Moral Authority of Nature*. Chicago: The Chicago University Press 2004, pp. 282-307; Charlotte Sleigh, *Ant*. Chicago: The University of Chicago Press 2003; John F. M. Clark, "A Little People but Exceedingly Wise? Taming the Ant and the Savage in Nineteeth-Century England", in: *La Lettre de la Maison Française,* Oxford, VII, 1997, pp. 65-83.

phasized the philosophical aspects of their science as well as the practical ones in order to show its usefulness.

4.3 On the origin of inequality among social insects

The main political themes concerning the insect societies are presented with a strong dramatization in *L'Insecte,* by Jules Michelet. The scholars who studied the work of Michelet took his naturalist books into account.[19] Michelet observed in his garden a "civil war" between large ants and small ones and was horrified by the cruel revenge of the small ones. But the greatest moral problem is slavery. In Michelet's eyes, rearing aphids is fair: it is like cattle breeding. But, the problem is different when ants capture the pupae of another species of ants, and carry them to their own nest, where the new born ants will work for them during their whole lives. This phenomenon of social parasitism was discovered by Pierre Huber, and named by him *slavery*.[20] Pierre Huber was the son of François Huber.[21] To put it in a nutshell, the father studied the bees and the son observed the ants. Though it is still used in entomological text-books, the word *slave* is questionable as far as the slaves and their masters belong to the same family – the formicidae – but not to the same species: the slave makers observed by Pierre Huber, are Amazon ants (*Polyergus rufescens*) while their so-called slaves belong to another species: ash-coloured ants (*Formica rufa*). In this case, such a taxonomical point of view is not taken into account, and all authors use the word *slave* without considering it metaphoric. If they agree on the name and the description of the phenomenon, authors differ in their moral judgment. Pierre Huber considers that these slaves have no memories of their motherland and he is convinced that they are happy in their new colony.[22] Michelet, on the opposite side, is depressed to see Nature setting a bad example of injustice and servitude.

> What! I turn aside from the history of men in search of innocence; I hope at least to discover among beasts the evenhanded justice of Nature, the primitive rectitude of the plan of Creation. I seek in this people whom I had previously loved and esteemed for their laboriousness and temperateness, the severe and touching image of republican virtue … and I find this indescribable horror![23]

19 Roland Barthes, *Michelet*. Paris: Seuil 1954; Linda Orr, *Jules Michelet, Nature, History and Language*. Ithaca: Cornell University Press 1976; Edward K. Kaplan, *Michelet's Poetic Vision. A Romantic Philosophy of Nature, Man, & Woman.* Amherst: University of Massachussets Press 1977; Georges Gusdorf, *Le Savoir romantique de la Nature*. Paris: Payot 1985.

20 Pierre Huber, *Recherches sur les mœurs des fourmis indigènes*. Paris et Genève: Paschoud 1810. (Translated into English in 1820 under the title *The Natural History of Ants*.)

21 See François Huber, *Nouvelles observations sur les Abeilles*. Genève: Barde, Manget 1792.

22 Pierre Huber, *Recherches, op. cit*., p. 210.

23 See Michelet, *L'Insecte*, *op. cit.,* pp. 259-260. English translation quoted in Kaplan, *op.*

4.4 Social insects and evolution

Depressed by this treason of nature, Michelet seeks some relief in an evolutionistic approach. He suggests that slave maker colonies are monstrous societies, deprived of the working part of the people. So nature, far from legitimating injustice, reveals it as degeneration. As Roland Barthes says in his essay: "Michelet does not naturalize morality, he moralizes nature."[24]

An opposite phrase can be used to epitomize the point of view expressed by Marcelin Berthelot, the famous French chemist and politician, who observed ants as a hobby. He published in 1886 under the title *Science et philosophie* a book collecting several essays. One of them is called "Les cités animales et leur évolution".[25] Berthelot stated that the subject was always in the mind of savants and philosophers, because of the analogies between animal societies and human ones. He was convinced that the same instinct of sociability was active among human races and among animal ones. He considered the hypothesis of the social contract as a chimerical one. Berthelot knew the classical objection of the stability of animal societies contrasting with the historical change undergone by human societies, but he dismissed it, opposing the vicissitudes of an ant colony, which he observed in a wood near Paris. Ten years later, in another collection of essays called *Science et morale*, Berthelot devoted a paper to ant invasions.[26] He considered that it is more useful to compare human societies with ant colonies than with beehives, because while in the latter laws are uniform, in the former there is a place for the individual initiatives. In 1903, Marcelin Berthelot, in a commentary on Michelet's *L'Insecte,* suggested that Michelet, in his books on natural history, searched the symbolism of his own thought.[27]

Charles Darwin was keenly interested in entomology.[28] An entire chapter of the *Origin of Species,* chapter 7 in the first edition, is devoted to instinct structure for "the welfare of each species, under its present conditions of life".[29] Among the examples of instincts analyzed by Darwin, "the slave-making instinct of certain ants" and "the comb-making power of the hive bee", are borrowed from the study of social insects.[30] Darwin proposed to explain them as complications of simpler instincts by natural selection. The ancestors of the slave-making ants might have

---

cit., p. 87.

24 Barthes, *Michelet, op. cit.,* p. 35.

25 Marcelin Berthelot, "Les cités animales et leur evolution", in: *Science et philosophie*. Paris: Calman-Lévy 1886, pp. 172-184. (I wish to thank Annie Petit for this reference.)

26 Marcelin Berthelot, "Les sociétés animales. Les invasions des fourmis; le potentiel moral", in: *Science et morale*. Paris: Calman-Lévy 1897, pp. 313-331. (I wish to thank Annie Petit for this reference.)

27 Marcelin Berthelot, "Etude. Lettre à monsieur Ludovic Halévy", in: Jules Michelet, *L'Insecte*. Paris: Calman-Lévy 1903, pp. 1-39.

28 Yves Carton, *Entomologie, Darwin et Darwinisme*. Paris: Hermann 2011.

29 Darwin, *On the Origin of Species, op. cit*., p. 209.

30 Darwin, *ibid.,* p. 216.

just displayed the instinct of stealing and storing the pupae of other species of ants as a source of food. Some of the pupae, which were stored, might have developed into workers. These workers might have followed "their proper instincts, and do what work they could".[31] Their work being useful, the natural selection gradually complicated the instinct of making raids on nets of other species of ants, and transformed this instinct into the habit of capturing workers. Concerning the comb-making power of the hive bee, the explanation rests upon a gradual process from humble-bees "using the old cocoons to hold honey", to the hive-bee, *Apis mellifera* constructing hexagonal prisms with bases made of three rhombs.[32] In this gradual process, the Mexican bee, *Melipona domestica*, with its cylindrical cells plays the role of an intermediate stage. Darwin reports on some experiments he performed with ants and bees. He also thanked several naturalists for their advice, and among them, a specialist of crystallography, William Miller (for the geometrical approach of the comb of the Bees).

The first issue in 1874 of *Les Fourmis de la Suisse* by Auguste Forel, the publication in 1877 of the doctoral dissertation of Alfred Espinas, *Les sociétés animales*, and the popular success of Maurice Maeterlinck's *Vie des Abeilles* – followed in 1926 by *La Vie des Termites,* and in 1930 by *La Vie des Fourmis* – are evidences of the importance of the Insect societies as a matter of reflection for scientists and philosophers in the second half of the 19th century and the beginning of the 20th century.[33] The celebrated *Mémoires entomologiques* of Jean-Henri Fabre, despite the few pages devoted to ants and bees, can be taken into account to assess the place of entomology in the early 1900.[34]

A contrast between the evolution of insects and the evolution of humans was suggested by Henri Bergson in *L'Evolution créatrice* (1907). Bergson presented the history of life as a road with two major bifurcations. The first one separates vegetables from animals, the second one separates arthropods from vertebrates. The first line, the arthropods' one, goes toward instinct, its climax being the Hymenoptera, which is the order of insects to which belong Ants and Bees. The second line, the vertebrates one, goes toward intelligence, its climax being man.[35] So, Bergson's view of evolution of life avoided the myth of a linear progress, as far as he took into account insect societies.

---

31 Darwin, *ibid.,* p. 223.

32 Darwin, *ibid.,* p. 225.

33 Auguste Forel, *Les Fourmis de la Suisse.* Bâle, Genève, Lyon: Georg 1874; Alfred Espinas, *Des Sociétés animales*, 2nd ed. Paris: Germer, Baillière et Cie, 1878. [Reprint: New-York: Arno Press 1977]; Maurice Maeterlinck, *La vie des Abeilles*. Paris: Fasquelle, 1901; *La vie des Termites*. Paris: Fasquelle 1926; *La vie des Fourmis*. Paris: Fasquelle 1930.

34 Fabre, *op. cit.* See also Fabre, *Souvenirs entomologiques*. Yves Delange (Ed.). Paris: Robert Laffont , 2 vol. 1989 (coll. Bouquins).

35 Henri Bergson, *L'évolution créatrice*. Paris: PUF 1907. New issue: 1962, ch. II, p. 135.

4.5 Sociobiology vs. swarm intelligence

A conflicting view on insects and human evolution occurred with the debate on socio-biology. The starting point looks like a strict question of biomathematics. In 1964, William Hamilton, a British biologist, published, in two issues of the *Journal of Theoretical Biology*, a paper entitled "The genetical evolution of social behaviour". Hamilton outlined a hypothesis connecting the social behaviour in the hymenoptera with the number of chromosomes. Among ants and bees, females are diploid, which means they have *2n* chromosomes like any animal, while the males are haploid, which means they only have *n* chromosomes. The theory of probabilities predicts that in this case, a female can have three quarters of her genes in common with any of her sisters, but only half of her genes with an offspring. So from the point of view of the genes, an ant must prefer nursing her sisters to having a daughter. Of course this can be applied also to the bees, which belong to the hymenoptera.[36] The difficulties arose when one postulated that every social behaviour in any animal species – including *Homo sapiens* – can be linked with a genetic fact in a similar manner. Edward Wilson, an entomologist and theoretician of scientific ecology, specialized in the study of ants, claimed for a larger extension of socio-biology. He did it in a provocative way stating, in 1976, that "the division between biology, particularly population biology, and social sciences, no longer exists".[37] Such a claim needs an epistemological discussion. Stated as a fact, it appeared, to social scientists, at least, like a threat of annexation of their discipline. The controversy took a political turn: a biological theory of social behaviour being *a priori* suspected of legitimating inequalities.

Wilson's view looks reductionist. Not surprisingly, a view, more in harmony with the cultural atmosphere of the seventies, has been launched. The works of Pierre-Paul Grassé (at the end of the 1950s) on the construction of a nest by the termites, a paper of Remy Chauvin (1974), opened a route to Jean-Louis Deneubourg and several other scholars.[38] The beehive or the ant colony offer striking examples of collective intelligence. An ant colony grouping a great number of individuals with very simple behaviour can solve complex problems, like finding the shortest way between several points: the salesman problem.

---

36 William Hamilton, "The Genetical Evolution of Social Behaviour", in: *Journal of Theoretical Biology*, 7, 1964, pp. 1-16 and pp. 17-52.

37 Edward O. Wilson, "The Central Problem of Socio-biology", in: Robert May (Ed.), *Theoretical Ecology: Principles and Applications*. Oxford: Blackwell, pp. 205-217 (quotation, p. 217).

38 Pierre-Paul Grassé, "La reconstruction du nid et les coordinations inter-individuelles chez *Belliocstermes natalensi* et *Cubitermes sp.*: la théorie de la stigmergie: essai d'interprétation des termites constructeurs", in: *Insectes sociaux*, 6, 1959, pp. 41-83; Rémy Chauvin, "Les sociétés les plus complexes chez les Insectes", *Communications*, 22, 1974, pp. 63-71; Jean-Louis Deneubourg et al., "The Dynamic of Collective Sorting. Robot-like Ants and Ant-like Robots", in: J. A. Meyer and S. Wilson (Eds.), *From Animals to Animats*. Cambridge (Mass.): The MIT Press, pp. 346-354.

5. Conclusion

First, Insects do not *look* small, they *are* small. Second, the works of the entomologists have resulted in a distinction between insects and crustaceans, arachnids, centipedes, as well as the definition of division in several orders of insects. This classification is a social construction, in the sense of the sociology of scientific knowledge, but it is a construction, which, far from being just more or less convenient, can be more or less realistic for theoretical reasons. Third, an ant colony or a beehive is a society, which can be thought of as analogous to a great city or to a brain. Entomology offers to philosophy many stimulating problems, which have not been fully exploited.[39]

Muséum National d'Histoire Naturelle
Centre Alexandre Koyré (EHESS, CNRS, MNHN)
57, rue Cuvier
75005, Paris
France
jmdrouin@wanadoo.fr

39 I wish to thank Jean-Jacques Levive and Frank Egerton for their careful reading of this text.

Anastasios Brenner and François Henn

Chemistry and French Philosophy of Science. A Comparison of Historical and Contemporary Views

Abstract

Philosophers of science have shown over the past several years a growing interest in chemistry. Chemistry has always held an important place in French philosophy of science. By confronting our respective experiences as philosopher and chemist, we bring out the specificity of the French tradition. The insight provided thereby will allow us to examine afresh some philosophical problems raised by contemporary science: changing conceptions of matter, laboratory practice as opposed to mathematical representation as well as the impact of computer modeling and atomic microscopy on our knowledge of the behavior of matter.

1. Introduction

Since the 1990s there has been a significant attempt to promote a reflection especially devoted to chemistry. One may point to the founding of an international society and two journals in this area: the International Society for the Philosophy of Chemistry (1997), *Hyle: International Journal for Philosophy of Chemistry* (1995), *Foundations of Chemistry* (1999). This movement is a reaction against the neglect of chemistry in mainstream philosophy of science. Current interest in this science may be viewed as a late consequence of the critique of logical empiricism, which has led to pay greater attention to concrete, practical or technical aspects of science.

Now chemistry has held an important place in French philosophy of science. When philosophy of science was just beginning to emerge in the 18th century, such prominent thinkers as Jean-Jacques Rousseau and Denis Diderot wrote some substantial texts on chemistry. Later, in 1835, Auguste Comte devoted five lessons of his *Cours de philosophie positive* to what he called "Chemical philosophy".[1] In the aftermath several French thinkers paid attention to this science: Pierre Duhem, Émile Meyerson, Gaston Bachelard, Hélène Metzger, and more recently François Dagognet, Bernadette Bensaude-Vincent, Isabelle Stengers. We shall focus on the former three, because they established contemporary philosophy of chemistry. In-

1 Auguste Comte, *Cours de philosophie positive* (1839–1842), 2 vols., Paris: Hermann 1998. Abridged translation H. Martineau, *The Positive Philosophy of Auguste Comte*, 2 vols., Cambridge: Cambridge University Press 2009.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_31,

deed, Duhem, Meyerson and Bachelard offered significant studies of the chemist's approach. And, by examining their conceptions, we can bring out some characteristic features of the French tradition.

We have chosen to take up our topic by drawing on our respective areas of competence as philosopher and chemist. Let us explain this choice.

1.1 A philosopher's motives

As a philosopher I am sensitive to the call for a broader perspective with respect to the sciences. The philosopher can no longer remain content with a formal analysis of science nor restrict herself to those fields that submit easily to such a task. She should henceforth take into account the diverse results of science studies (historical, sociological, philosophical, anthropological, etc.) as well as the numerous methods employed in the sciences.

In this respect chemistry raises questions that are of primary interest to the philosopher: the relations between the organic and the inorganic, the nature of the basic stuff of the universe, the effect of drugs on us, the promise or the fright raised by nanotechnology.

1.2 A scientist's motives

As a chemist I expect from the philosopher a better understanding of what distinguishes chemistry from the other natural sciences, those bordering on its domain: physics, biology, the earth sciences, etc. It is important to grasp the philosophy that underlies this science – a question that involves chemistry's relation to society. This quest for meaning is not without practical import, as it may provide clues to two problems:

(1) Why is chemistry so often relegated to the margins of the so-called exact sciences? Why is it invariably given a bad reputation – both by the lay public and by scientists – associating it with the "dark ages" of alchemy and magic? In particular, why is it that, as society becomes more critical of techno-science, chemistry is designated as the culprit, retaining from a bygone era a quasi-malefic image? This is a reputation which it cannot get rid of in spite of its success in areas such as the medical or environmental sciences, for example. And this holds regardless of the ever greater recourse to the methods of physics, which are well-established on the theoretical level. In other words, can we make clear the connections between chemistry – its ontology, its epistemology – and the other sciences as well as its connections with society?

(2) What are the consequences of recent discoveries with which physicists and computer scientists have provided the chemist? I am referring to certain innovations in microscopy, on the one hand, and to supercomputers, on the other. How are these going to modify the representations that first the chemist and then the lay person are coming to have of the atomic and molecular world? This is a fundamen-

tal issue, which is bound to have, in my opinion, an effect on the creativity of the chemist, our conception of matter and the way chemistry is taught.

In response to these questions, can we find clues in the philosophical reflections on chemistry for advancing our comprehension of current chemistry – a science which claims to handle atoms and molecules at the nanoscopic level?

2. Rousseau and Diderot: Chemistry as an epistemological model

Jean-Jacques Rousseau (1712–1778), following Georg Ernst Stahl, conceived of chemistry by analogy with mechanics and according to the general principles of attraction and repulsion among bodies as evidenced in his *Chemical Institutions* of 1747. At the same time, this science was clearly set apart, because chemistry "makes it possible to reach a true knowledge of nature, that is the bodies that compose it".[2] Chemistry was then as much a science that transforms bodies as one that separates them, since each stage of separation can be conceived as a transformation. Analysis, understood as the transformation of bodies, is carried out by a series of manipulations, that is literally manual operations, before becoming a mental procedure. Rousseau dwelled on the concepts of "mixt" and "aggregate"[3] which he distinguished precisely: in mixed bodies the original properties of a constituent are no longer recognizable. "Mixion" consists then in an alteration, whereas "aggregation" is a composition obeying a strict logic of immanence. This contrast became a paradigm for Rousseau's political concepts of association and body.[4] To sum up, as stated by Bruno Bernardi, Rousseau's approach is "a regional philosophy of chemistry"[5], which sets chemistry apart from the other natural sciences.

---

2 Jean-Jacques Rousseau, *Les institutions chimiques* (1747), Paris: Champion 2010, I, ch. 1, p. 10, our translation. Cf. Bruno Bernardi, "Pourquoi la chimie? Le cas Rousseau", in: *Revue Dix-huitième siècle* 42, 2010, pp. 433-443.

3 Following Louis-Bernard Guyton de Morveau, who wrote: "Aggregation is merely the union of several parts of a similar body without decomposition, and which are called in consequence integrated parts […]. Affinity, in contrast, makes up a new body", our translation. This passage from the *Encyclopédie* is quoted in Bernardi, *La fabrique des concepts*, Paris: Champion 2006, p. 158.

4 Rousseau: "The life of both bodies is the self common to the whole, the reciprocal sensibility and internal correspondence of all the parts. Where this communication ceases, where the formal unity disappears, and the contiguous parts belong to one another only by juxtaposition, the man is dead, or the State is dissolved", *Discours sur l'économie politique*, in: *Oeuvres complètes*, vol. 3, Paris: Gallimard 1964, p. 245. English translation, Constitution Society. Cf. Roger D. Masters and Christopher Kelly (Eds.), *The Collected Works of Rousseau*, vol. 3, Hanover, NH: University Press of New England 1993. See also Luc Vincenti, *Jean-Jacques Rousseau: l'individu et la République*, Paris: Kimé 2001, ch. 5, p. 146.

5 Bernardi, "Pourquoi la chimie?".

Denis Diderot (1713–1784) can also be considered as one of the first philosophers of chemistry. As emphasized by François Pépin[6], he added to his transcription of Guillaume-François Rouelle's lectures[7] a lengthy introduction, in which he situated chemistry – namely its scientificity and philosophical dignity – in historical perspective. Diderot viewed chemistry as a science distinguished not so much by its subject matter but rather its methods; these are formally very different from those of the mechanist. He sought out its specificities, its particular forms of activity – above all laboratory practice, which Diderot contrasts with the activity of mechanics. The latter he classed among the a priori speculations pursued in logico-deductive demonstrations. The methods of chemistry are to be understood in the light of its relations to nature, whose secrets it seeks to reveal by transforming it. These particularities are illuminating:

> You may devote yourself as much as you like to geometry and metaphysics, but I, who am a physicist and a chemist, I study bodies in nature and not in my mind, I see them existing, varied, bearing properties and actions, and moving about in the universe as in the laboratory, where if a spark goes off near three molecules comprising saltpeter, coal and sulfur an explosion will necessarily ensue.[8]

Rousseau and Diderot provide us with examples of a genuine endeavour to understand chemistry not only in terms of its usual objects of study but also its special operations and practices. What we have here is undoubtedly the elaboration of a philosophy of chemistry, a discourse on an area of scientific activity that takes into account its particularities. One notes the distinctive status that chemistry held in the minds of French philosophers of the 18th century. This science likewise played an important role in Kant, who made use of arguments similar to those of Rousseau and Diderot. He conceived of experimental chemistry as more an art than a science and moreover as an analogy for his transcendental idealism :

> So long therefore, as there is still for chemical actions of matter on one another no concept to be discovered that can be constructed, that is, no law of the approach or withdrawal of the parts of matter can be specified according to which, perhaps in proportion to their density

6 François Pépin, "Diderot philosophe de la chimie: des Lumières à la science contemporaine", in: Jean-Pierre Llored (Ed.), *La chimie, cette inconnue?* Paris: Hermann, in press, our translation.

7 Guillaume-François Rouelle (1703–1770) is one of the major French chemists of the 18th century. Many renown members of the French intellectual community, including Lavoisier, attended his lectures on chemistry, which he held in his laboratory. He was associate member of the French Academy of Science. Diderot followed his courses for three years.

8 Denis Diderot, *Principes philosophiques sur la matière et le mouvement* (1770), in: *Oeuvres philosophiques*. Paris: Garnier 1964, p. 395, our translation. See Jean Starobinski, *Action and Reaction*. New York: Zone Books 2003, ch. 2. and Pépin "Diderot: la chimie comme modèle d'une philosophie expérimentale", in: *Revue du Dix-huitième siècle* 42, 2010, pp. 445-472.

or the like, their motions and all the consequences thereof can be made intuitive and presented a priori in space (a demand that will only with great difficulty ever be fulfilled), then chemistry can be nothing more than a systematic art or experimental doctrine, but never a proper science.[9]

3. Duhem: Physical chemistry, axiomatics and the Vienna Circle

Let us now turn to the situation at the beginning of the 20th century. Pierre Duhem (1861–1916), who offered an influential conception in philosophy of science, devoted particular attention to chemistry. In 1902 he published *Mixture and Chemical Combination*, in which he explored the historical development and the philosophical implications of this field.[10] Several other of his works touch on chemistry. This is not surprising as the research program in which Duhem was involved consisted in closing the gap between physics and chemistry, in order to develop a general thermodynamics or energetics, which was supposed to include chemical phenomena. One needs only to read Duhem's major work *The Aim and Structure of Physical Theory* in order to recognize that chemistry is not in the least forgotten.[11] It receives no less attention than electricity, magnetism or optics. Chemistry is mentioned at decisive moments in Duhem's argumentation: the concept of natural classification, the principle of simplicity and the holist thesis.[12]

Of course, Duhem's philosophy of chemistry arose in a particular context. He promoted the application of thermodynamics and in particular the concept of thermodynamic potential in the wake of Helmholtz and Gibbs. Yet Duhem remained hostile to atomism. Indeed, at the time, the multiplication of hypotheses and disagreements between theory and experiment raised much perplexity from the scientific community. To be sure, Duhem's chemistry is that of a physicist. He relied mainly on the tools of thermodynamics. His presentation was modeled on physics and was highly mathematical in style. What we have here, however, is not an attempt at reduction: Duhem provided an example of a unification that maintains the autonomy of the fundamental branches of science. It is important to note that Duhem's axiomatics left room for applied science. As Louis de Broglie remarked in the foreword to the English translation of *The Aim and Structure*: "It

9 Immanuel Kant, *Metaphysical Foundations of Natural Science*, translation M. Friedman, Cambridge: Cambridge University Press 2004, p. 7. See Mai Lequan, *La chimie selon Kant*. Paris: Presses universitaires de France 2000.

10 Pierre Duhem, *Le mixte et la combinaison chimique*. Paris: Fayard 1985. English translation P. Needham, *Mixture and Chemical Combination*. Dordrecht: Kluwer 2002. Cf. Robert Deltete and Anastasios Brenner, "Essay Review of Pierre Duhem's *Mixture*", in: *Foundations of Chemistry* 6, 2004, pp. 203-232.

11 Duhem, *La théorie physique, son objet et sa structure* (1906), Paris: Vrin 1981. English translation P.P. Wiener, *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press 1991.

12 *Ibid*., pp. 25, 127, 214.

is fair to insist on the fact that Duhem, though he was constantly preoccupied with the establishment of an impeccable axiomatic system in the theories he developed, never lost sight of the problems of applications. Notably in the domain of physical chemistry."[13]

The logical positivists or empiricists retained only axiomatics, leaving aside a large portion of Duhem's work. They devoted their attention to those fields to which logical analysis applies best: mathematics and selected parts of physics. The Vienna Circle's manifesto outlined a program which covers many "fields of problems": arithmetic, physics, geometry, biology, psychology and the social sciences (including history and political economy).[14] Chemistry is altogether missing. Although this science is obviously considered rigorous, it does not lend itself easily to an axiomatic treatment. The subsequent development of logical positivism in the *Encyclopedia of Unified Science* gives no reason to change this conclusion; the nineteen volumes of this series, founded by Otto Neurath, which appeared between 1938 and 1970, deal with a large number of sciences, but one does not find anything on chemistry.[15] That there was no chemist among the twenty or so members of the Vienna Circle does not furnish an explanation. Reading over the writings of the logical positivists, one also finds few references to chemistry. Let us quote one passage, however, if only to show that there was no intention of excluding this field from their "scientific worldview". Otto Neurath in the following passage called on chemistry to mark a difference between a genuine scientific inquiry and a mere logical study:

> If we investigate whether the directions of a chemist are in logical agreement, we are not yet doing work in chemistry. In order to be able to do work in chemistry, we must state certain correlations between certain chemical processes and certain temperatures and the like.[16]

Here as elsewhere Neurath criticized his fellow positivists for not paying heed to practice. But let us note especially the slight space allotted to this science. It is surprising that empiricists were not more attentive to the important experimental work accomplished by chemists, as regards both their findings and the concrete procedures they elaborated. It is furthermore strange that self-proclaimed materialists were not more interested in the problem of matter with which chemistry is inevitably involved.

---

13 *Ibid*., p. VI.

14 Hans Hahn, Rudolph Carnap and Otto Neurath, "The Scientific Conception of the World: The Vienna Circle", in: Otto Neurath, *Empiricism and Sociology*. Dordrecht: Reidel 1973, pp. 299-318.

15 Neurath provides a presentation of the project in "Unified Science and its Encyclopedia", in: *Philosophical Papers*. Dordrecht: Reidel 1983. Cf. Friedrich Stadler, *The Vienna Circle: Studies in the Origins, Development and Influence of Logical Empiricism*. Springer: Vienna 2001.

16 Neurath, "Sociology in the Framework of Physicalism", in: *Philosophical Papers*, p. 80.

4. Meyerson: The critique of positivism and the defense of atomism

The importance given by Duhem to chemistry would not have exerted an influence, if several major thinkers had not taken up the issues he raised. It is undoubtedly on account of the contributions of Meyerson and Bachelard that chemistry came to acquire its particular place in contemporary French philosophy of science. Émile Meyerson (1859–1933) was trained as a chemist. He was instrumental in spreading the methods and philosophical discussions he had acquired during his studies in Germany. Although Duhem's senior by a few years, he entered the philosophical debate later, his first book *Identity and Reality* being published in 1908.[17] One is not surprised then to see him go to pains to define his position in contrast to Duhem. He vigorously defended atomism as reassessed in the light of recent discoveries. He rejected the positivism of Auguste Comte and Ernst Mach as well as its reformulation by Duhem, and he asserted his own realism. Meyerson offered a different reading of history:

> Neither Galileo nor Descartes are atomists in the strict sense of the term, and amongst later physicists, many, even those who assert that they adhere to this doctrine, formulate its principles in a very different way, often with inexactness, and, in practice, frequently straying from it. Yet it is certain that one can unite all the conceptions under the term *mechanical* [*conceptions mecanistes*] and that their common characteristics are of considerable importance. On the whole, it does not seem too bold to affirm that the mechanical hypotheses had their birth with science and have formed one body with it during all the epochs in which it has really advanced.[18]

Rather than "mechanicist conceptions" some historians today prefer to speak of "kinetic corpuscularianism".[19]

Meyerson did not follow Duhem and Poincaré in conceiving scientific laws as freely chosen postulates. Rather laws are to be understood as the outcome of a struggle opposing the mind's tendency toward identity and the resistance of matter. As Meyerson wrote: "The activity of the mind does not appear to be completely arbitrary, but rather to be guided, on the one hand, by the identification of diversity and, on the other, by the observation of the behavior of concrete reality [*du réel concret*]."[20] Such descriptions are undoubtedly inspired by the practice of the chemist. And Meyerson compared, on more than one occasion, the activity of the scientist to the act of drawing a fiber from the magma of reality. To be sure this

17 Émile Meyerson, *Identité et réalité*. Paris: Vrin 1951. English trans. K. Loewenberg, *Identity and Reality*. London: G. Allan & Unwin 1930.

18 *Ibid.*, p. 88.

19 For example H. Floris Cohen, *How Modern Science Came into the World*. Amsterdam: Amsterdam University Press 2010, p. 221.

20 Meyerson, *Du cheminement de la pensée* (1931), 3 vols., Paris: Alcan, p. 613, our translation.

metaphor is borrowed from the biologist Arthur Balfour.[21] But it applies relevantly to all sciences dealing with complexity.

Meyerson represented the beginning of a shift away from positivism. This was continued by Bachelard, Koyré and their disciples. Positivism was receding in France precisely when positivism was gaining strength in Austria. Meyerson wrote, after becoming acquainted with the doctrine of the Vienna Circle toward the end of his life:

> I deem it useful to summarize here these conceptions as a whole (which appear to enjoy at present a certain success in German-speaking lands) precisely because, on a number of essential points, they go entirely against the views that I have put forth in my earlier writings and that represent the core of this book.[22].

5. Bachelard: Chemistry and applied materialism

It is well known that Gaston Bachelard (1884–1962) distinguished himself more and more clearly as the years went by from Meyerson. His criticism was directed as much at the interpretation of scientific results as their consequences with regard to scientific rationality. This does not mean that Bachelard returned to earlier conceptions; he subscribed to Meyerson's criticism of idealism and positivism as well as his defense of atomism and realism. But he went on to set himself apart from his predecessor and reformulated these positions.

Let us recall a passage of *The Rationalist Activity of Contemporary Physics*, which aims to bring out the consequences of the "new theory of quanta".[23] According to Bachelard, Meyerson's realism is as questionable as the positivism he opposed:

> One is ill-prepared to follow the evolution of modern atomistics if one accepts Meyerson's phrase stating that the atom is merely "strictly speaking a portion of space". This is a ready-made phrase [*formule réponse*] […] which treats lightly the huge problem of modern atomistics. It leaves aside too hastily the prudent restrictions of a positivist mind.[24]

And Bachelard went on to formulate his position: "Corpuscles border on invention and discovery, precisely in the realm in which we believe applied rationalism to be active."[25] If Bachelard claimed to be realist, rationalist and even materialist, it should be understood that his is a constructive realism, an applied rationalism and a rational materialism. He sought to open a new perspective thereby. His scientific

21 *Ibid.*, p. 138. Cf. p. 246.

22 *Ibid.*, p. 790, our translation.

23 Gaston Bachelard, *L'activité rationaliste de la physique contemporaine*. Paris: Presses universitaires de France 1951.

24 *Ibid.*, p. 86, our translation. Cf. Meyerson, *Identity and Reality*, ch. 9.

25 Bachelard, *L'activité rationaliste*, p. 87.

realism has nothing to do with common sense realism, and Bachelard went as far as to speak in the same passage of a "relativism of ontology". As is known, W. V. O. Quine will also transform Meyerson's realism, but in a quite different context.[26] The scientist makes ontological engagements, but one should understand in what sense: he does not hesitate to continuously adjust his ontology to scientific progress.

Bachelard was led to elaborate what he called "phenomeno-technics". Instruments, experimental setups or better laboratories make it possible to create new phenomena, which although artificial can be more significant than natural processes. What is at issue here is an experimental realism which puts emphasis on the concrete context whereby proof is produced. Bachelard then formulated a philosophical conception based on his experience as a school teacher in physics and chemistry at the beginning of his professional career, a conception that was at odds with that of logical positivism.

6. Conclusions

We may now reflect on the position of contemporary chemistry on the basis of these historical elements.

6.1 The relation to matter

Following a hardline positivism, the chemist refrains, at least professionally, from the question of what matter is in itself. Her focus is on the formulation of concepts dealing with understanding chemical reactions and analyses, that is the mechanisms of decomposition and recombination of matter. The major issues are "What is the result of the reaction of A on B?" and "Why does this occur?" or conversely "How to obtain C from certain synthons?"[27] They can also be understood in the light of the recent development of Quantitative Structure-Activity Relationship (QSAR).[28]

The very nature of matter or its essence is not on the agenda. The chemist accepts: first, that matter is made up of atoms which bond together to constitute

26 W. V. O. Quine, *From a Logical Point of View*. Cambridge (Mass.): Harvard University Press 1980, p. 45. For an attempt to combine instruments of analytic philosophy of science and historical epistemology, see Ian Hacking, *Historical Ontology*. Cambridge (Mass.): Harvard University Press 2002.

27 "A synthon is defined as a structural unit within a molecule which is related to a possible synthetic operation, the term was coined in 1967 by E. J. Corey", in Wikipedia.

28 "The basic assumption of all molecule-based hypotheses is that similar molecules have similar activities. This principle is also called Structure-Activity Relationship (SAR). The underlying problem is therefore how to define a small difference on a molecular level, since each kind of activity, e.g. reaction ability, biotransformation ability, solubility, target activity, and so on, might depend on another difference", in Wikipedia.

complex molecular systems and solids, and secondly, that the chemical or physical properties of matter are dependent on the nature of the atoms involved, their spatial arrangements and their motions. She restricts her "deep" exploration of matter to the atoms or more precisely to the electrons of their outermost shells, the so-called valency electrons which are responsible for chemical reactivity.[29]

Bringing together the major issues and paradigms mentioned above with regard to "chemical" matter makes it possible to reformulate the problem that concerns daily the contemporary chemist, that is "What are the relations between composition, structure and the reactivity of atomic constructions?" This question, of a very general nature and common to all forms of chemistry, whether organic, inorganic or hybrid, arises in particular in the case of supra-molecular structures such as proteins, in which molecular self-organization plays an essential role in regard to their functionality.

6.2 On representations, micro-vision and macro-vision

Although physicists and chemists fully agree with regard to the issues and paradigms mentioned, their methodologies or ways of comprehending matter are fundamentally different. At the molecular level the chemist "prefers" to represent directly atomic and molecular constructions, by likening atoms to balls and chemical bonds to solid lines. She builds her reasoning on pictures or figurative representations. The physicist proceeds rather by speculating on reciprocal space, which is associated with energy, and she schematizes matter and its properties with the help of abstract representations directly related to mathematical descriptions.

The origin and import of these differences is essential, because recourse to figurative representations limits the scope of thought to the space defined by the picture which is used. The chemist consequently tends to represent atomic reality in virtue of the paradigm mentioned above by restricting herself to the molecular level. This apprehension of reality may explain why she is suspicious of or even hostile to the complete mathematization of chemical reactions, which, according to her, cannot account for all complexities and subtleties inherent in matter. This characteristic inability of physics to account for chemical reality could be explained, according to her, by the fact that the concrete system involves a large number of atoms, which corresponds to a high degree of freedom.

As the chemist moves up to higher scales, those pertaining to the physical continuum, the media that are conceptualized without recourse to atoms (dimensions above the nanometer), she is obliged to change radically her approach, because she has no other choice in this case but to call on oversimplified abstract mathematico-physical descriptions on the one hand or on empirical knowledge that cannot be mathematized on the other. The transition from discontinuity to

29 Valency electrons are those which are associated with the highest energy orbital and which are directly involved in chemical reactivity. For more details see the Frontier-Orbital Theory.

continuity in physics is not without difficulties. But whatever these may be, the tool that is used remains mathematics.

It is within this gap between the figurative representation of atoms or molecules and the mathematized representation of macroscopic media that the chemist can assert her specificity, her freedom vis-à-vis the injunction of the logico-deductive powers of mathematics. It is in this gap that the chemist expresses her "art". It seems that we return to the antagonisms between the physicists Duhem and Poincaré and the chemist Meyerson.

We could conclude on this issue by stating that the view of the physicist and that of the chemist continue to be at odds as they were in the 18th and 19th centuries, and that the chemist is assigned to this in-between zone, as it were trapped between a science and an art.

6.3 Recent developments

It is important to acknowledge that recent developments in "numerical" chemistry, which make it possible to simulate structures, atomic trajectories, physical properties and even chemical reactivity of larger and larger sets of atoms could change irreversibly the relation between the model and the representation of reality.

Two factors more than any others appear to be operating in this evolution which could result in a profound mutation. On the one hand, the computer is conceived as a black box, which renders all but completely invisible the mathematics and the models with which it works. In other words, it has become a tool, which, in being used, requires no special ability in mathematical physics. It can be assimilated thus with the majority of analytic instruments which are used daily without any need to know their "deep" principles. On the other hand, the results obtained by numerical procedures are often translated into atomistic representations of matter; one can visualize the atoms and molecules in space, their movements, their rearrangements, etc. What we have is "a journey to the center of matter"! It is difficult to sort things out and to avoid being led to think that this representation is an entirely faithful photograph of reality. The virtual aspect of this representation is not necessarily scrutinized with care. This vision of reality mediated by the computer is furthermore enhanced by the simultaneous development of atomic microscopy, the results of which are equally translated into figurative representations.

This relation to matter relies first and foremost on the confidence placed on the instruments, because they are too sophisticated and too complex to be understood by the non-specialist. Secondly, the easy and readily available pictures tend to diminish the hostility that the chemist consistently shows vis-à-vis a strictly "mechanistic" approach to matter.

It is probable that Duhem, had he lived today, would refer to atoms. But what would he think of the uses made of these new technologies and especially the philosophical consequences of their use? The questions raised by Rousseau, Diderot, Duhem and Meyerson are still on the agenda. And Bachelard's objection

to the identification of the atom with a portion of space points to the complexities still inherent in our conceptions of matter. After the breakthrough of quantum mechanics at the beginning of the 20th century, it is obvious that the figurative representation of molecular or atomic systems in a "classical" 3D Cartesian space cannot be fully suitable. Yet it remains the most used picture of matter and its use is still very efficient.

Anastasios Brenner
Department of Philosophy
C.R.I.S.E.S.
Université Paul Valéry-Montpellier 3
route de Mende
34199, Montpellier
France
anastasios.brenner@wanadoo.fr

François Henn
Laboratoire Charles Coulomb
UMR 5221 CNRS
Université Montpellier 2
Place Eugène Bataillon
34095, Montpellier
France
francois.henn@univ-montp2.fr

Cristina Chimisso

The Life Sciences and French Philosophy of Science: Georges Canguilhem on Norms

Abstract

Although in the last decades philosophers have increasingly paid attention to the life sciences, traditionally physics has dominated general philosophy of science. Does a focus on the life sciences and medicine produce a different philosophy of science and indeed a different conception of knowledge? Here I present a case study focussed on Georges Canguilhem. Canguilhem continued the philosophical tradition of what we now call historical epistemology, and always referred very closely to the philosophy of Gaston Bachelard. However, whereas Bachelard primarily studied the history of chemistry and physics, Canguilhem turned to the life sciences, medicine and psychiatry. I shall argue that some crucial differences in how they regarded norms, an issue seldom emphasised by Canguilhem himself or indeed by his critics, stem from the sciences on which they concentrated.

1. Introduction

Philosophers have traditionally regarded science as the exemplar of knowledge. What they have had in mind at least since Kant, however, has often been physics, rather than the sciences in general. The life sciences in particular, especially if we include as I shall do here the medical sciences, seem to have less to offer to the ambitious philosopher who aims at a universal model of knowledge as a perfectly rational enterprise, regulated by an unchanging method. Life appears to resist the demands of reason. As Georges Canguilhem put it 'reason is as regular as an accountant, life is as anarchic as an artist'.[1] Medicine has also a rather ambiguous status, and many would deny that it is a science. Its generalizations always find severe limitations, and human beings, with their unpredictability, often prevent the formation of the clear picture that some philosophers would like to obtain.

In France, the legacy of Cartesianism did not encourage the philosophy of the life sciences, both for its mechanistic approach and for its search for 'clear and distinct' ideas. Nevertheless, and perhaps as a reaction to mainstream philosophy, the philosophy of the life sciences and of medicine, though as a minority interest,

1 Georges Canguilhem, "Note sur la situation faite en France à la philosophie biologique", in: *Revue de métaphysique et de morale* (1947), pp. 322-332; p. 326.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_32, © Springer Science+Business Media Dordrecht 2013

has blossomed in France at least since the nineteenth century. Has the focus on the life sciences produced a different philosophy of science and indeed a different conception of knowledge? This question of the impact of the science of reference on epistemology and beyond is at the core of this article. It cannot be investigated in full in this limited space, and indeed I believe that it is to be answered differently depending on the particular issues at stake. Rather, I shall present a case study focussed on Georges Canguilhem, who occupies a very central position in the history of French philosophy of the life sciences and medicine.

Canguilhem always referred very closely to the philosophy of Gaston Bachelard, and with the latter is the main representative of the philosophical tradition that we now call historical epistemology. Indeed, as he put it, it 'hardly needs saying' that his close connection of epistemology and history derived from Bachelard's 'teachings'.[2] Both Bachelard and Canguilhem always considered science in its historical development,[3] although there are some differences in the role that history of science plays in their works. Indeed, it has been famously said that while Bachelard's work is best characterised as historical epistemology, a more fitting description of Canguilhem's is epistemological history.[4] It is true that Bachelard did not produce a historical book comparable with Canguilhem's *La formation du concept de réflexe aux xvii^e^ et xviii^e^ siècles*, and that his interest in history was motivated by his epistemological concern. However, Canguilhem himself believed that in addition to being a philosopher, Bachelard was a historian if by historian of science one means the act of revealing the process of the edification of knowledge, that is to say, if by history we mean epistemological history. On the other hand, in my view Canguilhem is not less of a philosopher than Bachelard, and his *Le normal et le pathologique* is a thoroughly philosophical book.[5]

---

2 Canguilhem, "Le rôle de l'épistemologie dans l'historiographie scientifique contemporaine", in: Georges Canguilhem, *Idéologie et rationalité dans l'histoire des sciences de la vie.* Paris: Vrin, 1993 [1977], pp. 11-29, p. 20.

3 For an excellent exposition of the role of history in Bachelard's philosophy, see Georges Canguilhem, "L'Histoire des Sciences dans l'œuvre épistémologique de Gaston Bachelard", in: *Annales de l'Université de Paris* 33, 1 (1963), pp. 24-39; reprinted in Canguilhem, *Etudes d'histoire et de philosophie des sciences concernant les vivants et la vie* (Paris, Vrin, 1994 [1968]).

4 For this distinction, see Dominique Lecourt, *Marxism and Epistemology. Bachelard, Canguilhem and Foucault.* London: NLB, 1975, p. 166; Jean Gayon, "The Concept of Individuality in Canguilhem's Philosophy of Biology", in: *Journal of the History of Biology* 31 (1998), pp. 205-325; p. 307, n.8; Hans-Jörg Rheinberger, "Reassessing the Historical Epistemology of Georges Canguilhem", in: Gary Gutting (Ed.), *Continental Philosophy of Science*. Oxford: Blackwell 2005, pp. 187-197. Michel Foucault applied the label of 'epistemological history' to both Bachelard and Canguilhem: Michel Foucault, *The Archaelogy of Knowledge.* London: Tavistock, 1972 [1969], p. 190.

5 Canguilhem, "L'Histoire des Sciences dans l'œuvre épistémologique de Gaston Bachelard".

I will not linger on the distinction between historical epistemology and epistemological history. What is important here is that for both of them what counts as knowledge is historical. For Bachelard the mind – or the way we think – changes in time: he dedicated books to the emergence of the scientific mind, and indeed of the 'new' scientific mind, that describe how our knowledge of the world takes different forms in different periods.[6] Epistemology, as a consequence, can only be historical. Canguilhem particularly focused on concepts, and investigated them in their historical development; he also examined how their variations impact on our way of regarding ourselves and our environment, as in the case of the concepts of the normal and pathological.[7]

In the context of this strong continuity between Bachelard's and Canguilhem's philosophies, it is interesting to investigate whether they may have been taken along different paths by their respective benchmark sciences. Whereas Bachelard focused on chemistry and physics, Canguilhem focused on the life sciences, medicine and psychiatry. Their different focuses were rooted in their respective training: before studying philosophy, Bachelard obtained a degree in science, Canguilhem in medicine. In order to investigate whether their sciences of reference produced differences in their philosophies, I shall explore one particular aspect of their philosophies, namely their approach to norms. I shall not argue that all the differences between Bachelard's and Canguilhem's philosophies are to be attributed to their sciences of reference. This would amount to overlooking their philosophical originality. However, I do think that the roots of some of their differences can be found in the sciences they studied. Their respective approaches to norms will show this.

1.1 Norms and the sciences

Some of the norms at the core of Bachelard's and Canguilhem's respective works appear not to be linked to the observation of a particular science, but rather to a general epistemological and historiographical outlook. An example is the epistemological norms that they both employed in order to judge the scientificity of theories and practices. Bachelard's view of knowledge is strictly normative: he saw an epistemological break between 'common' knowledge and scientific knowledge, and judged theories and practices as either scientific or non-scientific by using current science as his norm. It is current science that dictates what counts as

6 Gaston Bachelard, *Le nouvel esprit scientifique.* Paris: Presses universitaires de France, 1991 [1934] (Engl. tr. *The New Scientific Spirit*. Boston: Beacon Press 1984); Gaston Bachelard, *La formation de l'esprit scientifique: contribution à une psychanalyse de la connaissance objective.* Paris: Vrin, 1993 [1938] (Engl. tr. *The Formation of the Scientific Mind*. Manchester: Clinamen Press 2002).

7 Canguilhem, *Le normal et le pathologique.* Paris: Presses Universitaires de France, 1999 [1966] (Engl. tr. Canguilhem, *The Normal and the Pathological* (New York: Zone Books 1989 [1966]); Georges Canguilhem, *La formation du concept de réflexe aux xvii^e^ et xviii^e^ siècles.* Paris: Presses Universitaires de France 1955.

scientific knowledge: the epistemological norm is therefore also a historical norm, and epistemology is as a consequence historical.[8] Canguilhem adopted Bachelard's normative approach and used it in order to construct historical narratives, notably that of the concept of reflex.[9] I have discussed elsewhere the specific applications of Bachelard's and Canguilhem's normative approaches to history, as well as their differences.[10] Here, however, I do not aim to look at general epistemological and historiographical norms whose application is not connected with the specificity of a particular science. Rather, I shall investigate norms that for Bachelard and Canguilhem operate within their particular sciences. This will show whether Canguilhem's view of norms differs in any significant way from Bachelard's *because* they investigated norms in different fields. Norms are important terms of comparison because both Bachelard and Canguilhem saw them as crucial in the practice and indeed theory of science. Bachelard lamented that Henri Poincaré could not recognize the normative character of science, and rejected the latter's distinction between scientific and moral activity, based on a distinction between facts and norms.[11] As far as Canguilhem was concerned, in medicine the norms that make us judge a state as healthy or pathological are at the very core of medical science and practice.

But if Bachelard discussed norms in chemistry and physics, whereas Canguilhem discussed norms in medicine, is any comparison possible? In other words, is it not obvious that their respective representations of norms were different because

8 Bachelard's normative view of science permeates the whole of his epistemology. However, some of his works are more explicitly set out to defend his normative view, notably: Bachelard, *La formation de l'esprit scientifique*; Bachelard, *La philosophie du non. Essai d'une philosophie du nouvel esprit scientifique.* Paris: Presses Universitaires de France 1988 [1940] (Engl. tr. Bachelard, *The Philosophy of No: A Philosophy of the New Scientific Mind.* New York: Orion Press 1968; Bachelard, *Le nouvel esprit scientifique*; Bachelard, *Le matérialisme rationnel.* Paris: Presses Universitaires de France 1972 [1953], especially the last chapter.

9 Canguilhem, *La formation du concept de réflexe aux xvii^e^ et xviii^e^ siècles.*

10 Cristina Chimisso, "The Tribunal of Philosophy and its Norms: History and Philosophy in Georges Canguilhem's Historical Epistemology", in: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 34, 2 (2003) pp. 297-327; Chimisso, *Writing the History of the Mind: Philosophy and Science in France, 1900 to 1960s.* Aldershot: Ashgate 2008; Chimisso, *Gaston Bachelard: Critic of Science and the Imagination.* London: Routledge 2001.

11 Gaston Bachelard, "Valeur moral de la culture scientifique", in: Didier Gil (Ed.), *Bachelard et la culture scientifique*. Paris: Presses Universitaires de France 1993 [1934]. Bachelard in particular lamented that a mathematician like Poincaré could not see that science is normative. Regarding the normativity of mathematics, see Debru's discussion of Jacques Bouveresse's ideas; Debru writes: 'Mathematics are the best example of a creative process based on rules and able to create new mental objects which may be used as new rules etc.', Claude Debru, "The Concept of Normativity from Philosophy to Medicine: An Overview", in: *Medicine Studies* 3 (2011), pp. 1-7, p. 2.

they were discussing norms operating in different areas? In fact, neither Bachelard nor Canguilhem regarded their examination of their respective sciences as an exercise confined just to the understanding of those sciences in question. Indeed, their examination of science was a way to learn general philosophical lessons, which was not restricted to the philosophy of science, let alone the philosophy of a particular science. Bachelard wrote not only that 'science in effect creates philosophy',[12] but even that 'reason should obey science, the most highly evolved science, science in the process of evolution'.[13] Particular sciences for both of these philosophers have something to teach philosophy. Chemistry for Bachelard should revolutionise philosophy, and lead to the destruction of metaphysics in favour of 'metachemistry'.[14] Canguilhem on his part declared that his study of normal and pathological was 'an effort to integrate some of the methods and attainments of medicine into philosophical speculation'.[15] In fact, the norms of particular sciences for them impact not only on philosophy, but indeed on how human beings should live. As I shall discuss in the next section, the norms that Bachelard observed in the practice of chemical and physics research were for him the norms that should governed our lives, both as ethical and social norms. For Canguilhem, the norms of medicine and psychiatry are aimed at determining what a normal and healthy human being is. This is why there are important aspects of their respective discussions of norms that make a comparison between their respective use of norms not only possible, but indeed crucial, both in order to understand their respective philosophies, and to investigate my initial question, namely whether a focus on different sciences produces different philosophical results.

2. Bachelard: Scientific, moral and social norms

Bachelard regarded science as a dialectic activity, indeed an activity that exhibits several types of dialectics. One type of dialectics for him takes place between the minds of the scientists: science cannot be carried out in isolation, because, according to Bachelard, it is dependent on mutual correction and indeed mutual surveillance among scientists. Interaction and correction, or rectification, ensure in Bachelard's view that individual scientists are not carried astray by their own imagination, desires and instincts, which are in fact epistemological obstacles to be overcome. For him, objectivity can only emerge out of social exchanges.[16] Sci-

12 Bachelard, *Le nouvel esprit scientifique*, p. 7.

13 Bachelard, *The Philosophy of No*, p. 122 (original: Bachelard, *La philosophie du non*, p. 144).

14 Bachelard, *La philosophie du non*, p. 52.

15 Canguilhem, *The Normal and the Pathological*, p. 34; (original: Canguilhem, *Le normal et le pathologique*, p. 8).

16 Bachelard, *Le rationalisme appliqué.* Paris: Presses Universitaires de France 1986 [1949].

entific activity, however, is also based on the interaction between the mind and the object. The mind rationalizes and rectifies the object, and the reflection on the object in turn rectifies the mind.[17] Bachelard calls this process rectification, but also purification, especially when referring to the mind. This process of rationalization and purification enables the mind to overcome its own private desires and instincts and its selfish and self-centred attitude, and to become increasingly objective and rational.[18] Moreover, science develops through a continuous polemical relation with its own past: current knowledge, in order to advance, must 'say no', in his own words, to previous knowledge, not to deny it completely, but to revise it, even radically, but always dialectically.[19] This dialectic movement of history again brings about increased rationality and decreased subjectivity.

The relationships that take place in scientific activity are also the model for social relationships in general. Indeed, the reasons why Bachelard always championed science and scientific education are profoundly moral. Michel Serres famously declared that Bachelard's epistemological obstacles are in fact deadly sins, including sloth, lust, pride and covetousness.[20] Though polemically, Serres has stressed the profoundly moral message of Bachelard's philosophy of science. Scientific activity for Bachelard 'purifies' the mind from subjective desires and selfish attitudes. The scientific mind for Bachelard exhibits 'industriousness', that he contrasts with the 'laziness' of certain obsolete philosophical minds, which stubbornly refuse to follow the most advanced science. The new philosophy will follow science; in his words: '[Rationalism] is the consciousness of a rectified science, of a science which bears the mark of human action, of the well-considered, industrious, normalising action'.[21] Scientific activity, in short, is a way not only to achieve objectivity in the knowledge of nature, but also in our social relationships. For Bachelard science produces the norms not only of its own practice, but also of moral and social behaviour.

The model of science that Bachelard had in mind was mainly borrowed from chemistry and physics. The space of the laboratory is his model for the interactions that among scientists. Indeed, the metaphor of the scientific city that Bachelard employed shows that he regarded science as an activity that take place in designated and strictly public spaces, separated from everyday life and its concerns and emotions. Indeed, Bachelard illustrated the difference between chemistry and alchemy also in terms of space: chemistry, a science, is produced in public spaces, clearly distinct from the researchers' private dwellings, whereas alchemy was practiced in private spaces, where the authority of the master prevailed over

17 "Idéalisme discursive" in: Gaston Bachelard, *Etudes.* Paris: Vrin 1970; Bachelard, *Le rationalisme appliqué.*

18 Bachelard, *La formation de l'esprit scientifique.*

19 Bachelard, *La philosophie du non.*

20 Michel Serres, "La réforme et les sept péchés", in: *L'Arc* 42 (1970), pp. 14- 28.

21 Bachelard, *Le rationalisme appliqué*, p. 123.

rationality and objectivity.[22] The practice of chemistry for him also teaches us how to understand the role of analysis and synthesis in philosophy and more generally in our understanding of the world.[23] Similarly, subatomic physics for Bachelard shows to the philosophers the nature of rationalism, which is always a 'co-rationalism', as it springs out of objective interactions. Physics, and the manner in which it reconstructs its own history, for Bachelard indicated to us the *pedagogical* importance of what he called 'recurrent' history, a history that starts from the present and 'discovers, in the past, the progressive formation of the truth'.[24] What happens if we look at other sciences, which have different procedures and different social practices? What if, like Canguilhem, we concentrate on medicine and psychiatry? The model of the laboratory does not seem to be so central any more, and the practices of the medical sciences cannot be separated from every-day life, which indeed is the field in which they operate. Would norms, not to mention philosophy of science as a whole, change when confronted with the life sciences and medicine?

3. Canguilhem: Norms and normal human beings

Norms are at the very core of Canguilhem's work; indeed, one could say, as Claude Debru has done, that '[h]is philosophy is a commentary on the idea of norm'.[25] Unlike Bachelard's, however, his concept of norm derives from a reflection on the use of norms in medicine. Against Auguste Comte and Claude Bernard, he argued that a norm is not an average but a desired state of affairs. Canguilhem's examples are in the context of the life sciences and medicine. For instance, he argued that the dramatic increase in life expectancy that has taken place in the last century in Europe is due to norms of public hygiene and better living conditions. These norms (as states of affairs to which people aim to conform) have brought about what is then expressed as a number, the 'fact' of the average life expectancy. The norms of public hygiene are valued practices, and these created what we think of as an average. A normal life span, therefore, is not something that corresponds to an average, but it is an expression of norms as values, as desired state of affairs. In other words, norms exist because human beings are normative, that is to say they are able to produce new norms. As Canguilhem put it: "[i]f we can speak of the normal man as determined by the physiologist, it is because normative men exist for whom it is normal to break norms and establish new ones".[26]

22 *Ibid*., pp. 132-133, Bachelard, *La formation de l'esprit scientifique*, p. 50.

23 Bachelard, *Le matérialisme rationnel*, p. 147; Bachelard, *Le pluralisme cohérent de la chimie moderne.* Paris: Vrin 1973 [1932], Ch. 3.

24 Bachelard, *L'activité rationaliste de la physique contemporaine.* Paris: Presses Universitaires de France, 1951, p. 26.

25 Claude Debru, *Georges Canguilhem, science et non-science.* Paris: Editions rue d'Ulm / Presses de l'Ecole normale supérieure 2004, p. 83.

26 Canguilhem, *The Normal and the Pathological*, pp. 64-65 (original: Canguilhem, *Le*

Can Canguilhem's view of norms only be applied to medicine and the life sciences? I do not think so. In fact, it is possible to draw a parallel between Canguilhem and Bachelard, despite the vastly different contexts. For Bachelard norms are also desired states of affairs: scientific norms are produced in a continuous effort to make both the object and the mind more rational. Purification and the elimination of irregularities and of variations that are not relevant to a specific scientific aim for him not only give us true scientific theories and good techniques, but also morally desirable behaviour and models of social relationships. So far, it seems that their different sciences of reference have not brought about very significant differences between Bachelard's and Canguilhem's respective views of norms. However, differences are brought into relief if we ask ourselves who produces the norms we use. For Bachelard, scientific activity produces the norms. Scientific activity is an eminently human activity, as scientists produce not only knowledge, but also scientific objects and ultimately their own minds. However, science is not produced by all human beings, or in all places: he regarded the scientific community, the 'scientific city', as it called it, and the laboratory as particularly apt ideal and real spaces respectively for rational interactions. He was not, however, completely elitist, and also proposed schools, or at least ideal schools, as the model of human interactions based on objectivity, selflessness and rationality.[27]

Canguilhem, on the other hand, looked at the norms of health and pathology, both physical and mental. He did not think that these norms could be produced by experts, these being medics or psychiatrists. For him it is the patient who knows if she suffers, and if she needs to be restored to a previous normative state. Canguilhem argued that both the healthy and the pathological state are normative states, as they both obey norms. These norms, however, are not just different, but are also more or less rigid. For Canguilhem, a healthy person is more able than a sick one to adapt to a new environment. For the sick person, changes in the environment can be catastrophic, as he is not able to create new norms that would allow him to adapt to the new situation. As Daniel Lagache put it, illness is an inferior norm because it implies a partial loss of one's normative ability.[28] For Canguilhem norms are constantly produced, and they are produced at many levels. The most general level is that of life: Canguilhem believed that vital norms dictate the preservation of life. These norms are not only shared by human beings, but by living beings in general. If human – or other – beings did not have such norms, they simply would have not survived, and would be extinct. Human beings, however, do not only live by biological norms. Their difference with other living beings is precisely that they creatively produce their own norms, which can vary quite remarkably. Social and cultural norms are crucial for their lives, and for the way in which they judge their wellbeing. This also explains why norms of wellbeing are historical:

normal et le pathologique, p. 106).

27 Bachelard, *La formation de l'esprit scientifique*, p. 252.

28 Daniel Lagache, "Le normal et le pathologique d'après M. Georges Canguilhem", in: *Revue de métaphysique et de morale* 51, 4 (1946), pp. 355-370, p. 364.

because human beings continually produce them. Social and cultural norms shape the environment, and human beings themselves. They are central in determining what counts as normal. Canguilhem even conceded that cultural norms can be at odds with vital norms. For instance, he quoted Pascal writing that good health is a perilous state for human beings, as it is a danger for their souls; illness is the state in which human beings should spend their lives.[29] In this case, the value of eternal wellbeing clashes with life's principle of self-preservation. Pascal's values on the whole are foreign to modern societies. However, many habits in modern societies go directly against preserving our health, including smoking, excessive drinking or unhealthy diets. Here other social and cultural norms interfere with the norms of struggle against disease and of the preservation of life.

Canguilhem argued that each individual is in a unique position: no individual is the same as another, and indeed what counts as normal and as pathological is not the same for different people. It goes without saying that a twenty-year old will not judge her own normal state by the same norm as an eighty-year-old, and that the country in which an individual lives and his social class and occupation would have an impact on what counts as normal. But norms apply not only to groups, but also to individuals, not only because social norms interact differently in different individuals, but also because an individual's life-style, genetic make-up, medical history as well as her individual choices and individual reactions to her environment all play a role. Canguilhem puts the individual at the very centre of medical norms. The individual is at the cross-road of vital, social, cultural and individual norms, and in each case their interaction will be different.[30]

4. Conclusion

For Canguilhem the object of medicine is the normative individual, who is a full individual who cannot be split into a 'diurnal man' and a 'nocturnal man' as Bachelard proposed in his own philosophical anthropology.[31] Bachelard's diurnal man aims to suppress his emotions and imagination in the exercise of rationality, as emotions and the imagination continuously create epistemological obstacles that have to be overcome. Bachelard did not believe that the imagination, and our desires and instincts should be suppressed completely. In fact, he thought that they should be exercised, but in a rigorously private sphere, and on an individual basis.

29 Canguilhem, *Etudes d'histoire et de philosophie des sciences concernant les vivants et la vie*, p. 409.

30 On vital and social norms in Canguilhem, see Pierre Macherey, "Normes vitales et normes sociales dans l'*Essai sur quelques problèmes concernant le normale et le pathologique*", in: F. Bing, J-F. Braunstein, and Elisabeth Roudinesco (Eds.), *Actualité de Georges Canguilhem. Le normale et le pathologique*. Paris: Synthélabo 1988, pp. 71-84.

31 Bachelard, *Le matérialisme rationnel*, p. 19.

Emotions, dreams and desires are the realm of the 'nocturnal man', that it to say each individual's private sphere. In our private reverie, for him we can let our unconscious work, but not in the social sphere, where rationality should prevail. By contrast, in Canguilhem's view of the creation of norms, no separation of rationality and emotions can be drawn, or indeed should be drawn. François Dagognet has remarked that while Bachelard wanted an 'applied rationalism' and a 'rational materialism', Canguilhem proposed a 'rational vitalism'. He rightly added that the application of medical techniques does not command the organism, and does not impose its own directives to it, but rather aids the organism's self-development.[32]

For all the parallels between Bachelard and Canguilhem, this shows a real difference. The object of medicine cannot be rectified, or purified. Their suffering, hopes and dreams cannot be bracketed, indeed the norms of medicine are the result of these emotions. Canguilhem characterized the object of medicine as follows: "The sick person is a Subject, capable of expression, who recognizes himself as a subject in all that which he does not know how to designate other than with possessives: his pain and the representation that he makes of it, his angst, his hope and his dreams."[33] For Bachelard, human beings project emotions and dreams onto their objects (obviously the objects of chemistry and physics do not have emotions independently of human beings), but in order for these objects to become scientific, these emotions must be removed. The only 'mark' that these object should bear is that of rationality. As objects become rational, they become standardized: this can be seen as a reasonable aim for the objects of chemistry and physics, but it appears to be less so when we have to consider human beings. The objects of medicine, as they cannot be 'rectified', remain individuals with all their differences and irregularities. As Paul Rabinow has written, '[I]t is suffering, not normative measurements and standard deviations, that establishes the state of disease. Normativity begins with the living being, and with that being comes diversity'.[34] For Canguilhem the norms that are at the roots of medicine are not rational, or not only rational. This does not mean that rationality is not central to his project, and indeed to medicine. As Claude Debru has commented, for Canguilhem rationality plays a role of regulation of human activities.[35] Medicine is no exception. However, Canguilhem does not insist on any clear separation between rationality and emotions, and both are central, in their different ways, to medicine. The rationalization process can only operate within the norms that are rooted in

32 François Dagognet, "Une Oeuvre en trois temps", in: *Revue de métaphysique et de morale* 90, 1 (1985), p. 32.

33 Canguilhem, *Etudes d'histoire et de philosophie des sciences concernant les vivants et la vie*, p. 409.

34 Paul Rabinow, "Introduction: A Vital Rationalist", in: François Delaporte (Ed.), *A Vital Rationalist: Selected Writings from Georges Canguilhem*. New York: Zone Books 1994, pp. 11-22, p. 16. See also Guillaume Le Blanc, *Canguilhem et le normes.* Paris: Presses Universitaires de France 1998, pp. 62ff.

35 Debru, *Georges Canguilhem, science et non-science,* p. 83.

emotions. In fact, Canguilhem referred to Bachelard on norms and value, but he referred to the valorisation of imagination, and to his work on reverie.[36]

For both Bachelard and Canguilhem knowledge and its norms are historical. However, for Bachelard scientific knowledge is the realm of rationality, which he saw as historical; by contrast, the imagination, which gives science its epistemological obstacles, does not participate in the historical development of science and rationality. Bachelard regarded the unconscious, in which our images, instincts and desires are rooted, as essentially a-historical.[37] In Canguilhem we do not find this opposition between historical rationality and a-historical emotions. He believed that what counts as normal is historical. However, what counts as normal is not the result of rational investigation alone, but rather of a complex interaction of elements that include emotions and dreams, and indeed life. His science – or technique – of choice did not allow for a clear separation between rationality on the one hand, and emotions and dreams on the other. As a consequence, he recognized that all the components of medicine are historical: norms are produced by human beings, and are therefore subject to change, no matter how much rooted in life they are. Bachelard's and Canguilhem's respective focus on different sciences creates different points of view which are often downplayed in the name of the continuity of the tradition of historical epistemology. In fact, these differences exist, and they are applicable not only to the philosophy of their respective sciences, but also to general philosophy and indeed human life. This is because both of them regarded their respective sciences as showing how philosophy should develop, and how human beings should live.

Department of Philosophy
Arts Faculty
The Open University
MK7 6AA, Milton Keynes
UK
c.chimisso@open.ac.uk

36 Canguilhem, *Le normal et le pathologique*, pp. 176-177.

37 Bachelard, *La psychanalyse du feu*. Paris: Gallimard 1949 [1938], pp. 15-16; (Engl. tr. Bachelard, *The Psychoanalysis of Fire*, trans. Alan C. M. Ross, Boston: Beacon Press 1964 [1938]).

Massimo Ferrari

Neglected History: Giulio Preti, the Italian Philosophy of Science, and the Neo-Kantian Tradition

Abstract

This paper offers a brief survey of the influence of Kantian and neo-Kantian philosophy on Italian philosophy of science. At the core of the story stands Giulio Preti, a former disciple of Antonio Banfi who offered a sophisticated version of the relativized a priori embracing the historical knowledge too.

1. Introduction

The Kantian and neo-Kantian tradition proper had a very little influence on Italian philosophy of science during the 20th century. Although undoubtedly correct overall, such a statement requires supplementation by a further historical remark concerning the context of Italian philosophy. Since the very beginning of the last century, Italian philosophy was dominated by the idealistic and so called Neo-Hegelian trends promoted by Benedetto Croce and Giovanni Gentile. None of them had any interest in philosophy of science or in epistemology, yet their influence was not only a widespread one, but also lies at the origins of the "idealistic reaction against science" illustrated by Antonio Aliotta in his famous book of 1912.[1] By adherents of the other Italian currents of thought, such as neo-scholasticism, spiritualism or later existentialism, the philosophy of science flourishing in Europe and America in the first decades of 20th century was ignored.[2] After the First World War the older positivism which aimed at being a kind of scientific philosophy had run its course, its last representatives having turned their attention towards ethics and moral sciences instead. This philosophical situation characterized the Italian philosophical landscape in general at least until the period immediately before the

1 Antonio Aliotta, *La reazione idealistica contro la scienza*. Palermo: Optima 1912 (English translation: *The Idealistic Reaction against Science*, translated by Agnes McCaskill, London: Mac Millan 1914).

2 A very good survey is offered by Friedrich Stadler, "History of Philosophy of Science. From 'Wissenschaftslogik (Logic of Science)' to Philosophy of Science: Europe and America, 1930-1960", in: Theo Kuipers, Dov M. Gabbay, Paul Thagard and John Woods (Eds.), *Handbook of the Philosophy of Science: General Philosophy of Science – Focal Issues*. Amsterdam: Elsevier B.V. 2007, pp. 577-658.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_33, © Springer Science+Business Media Dordrecht 2013

outbreak of the Second World War, even though individual scholars such as Ludovico Geymomat, Eugenio Colorni, Giulio Preti, and others had already become acquainted with logical empiricism and contemporary philosophy of science since the late 1930s.

2. The Italian tradition of Kantian philosophy and the philosophy of science

The Kantian and neo-Kantian tradition represented, in this context, a very marginal philosophical current. To be sure, in the late 19th century Carlo Cantoni and Felice Tocco elaborated, to some extent, a neo-Kantian philosophy. Both published extensive and up-to-date interpretations of Kant's critical philosophy and were well informed about the contemporary neo-Kantianism in Germany.[3] Tocco, in particular, was also interested in Kant's metaphysical foundation of mathematical science and in a paper that appeared in *Kant-Studien* he offered one of the first analyses devoted to the *Opus postumum*.[4] But neither Cantoni nor Tocco were sensitive to the attempts made by the neo-Kantians in Germany in order to reformulate the transcendental philosophy in agreement with the new frontiers opened by contemporary logic, mathematics, and physics. Even the most important Italian philosophical review inspired by neo-Kantianism – the *Rivista Filosofica* founded and edited by Cantoni – was in fact scarcely attentive to the neo-Kantian debate flourishing in the early 20th century, and especially within the Marburg School, about the Kantian philosophy in its relationship to contemporary scientific knowledge.[5]

Even at the very beginning of the 20th century the philosophy of science was not totally absent from the Italian context and, more generally, some Italian philosophers dealing with science and epistemology were indeed engaged in the European debate.[6] It is perhaps superfluous to recall here that Giuseppe Peano,

3 See Carlo Cantoni *Emanuele Kant*, vol. I, *La filosofia teoretica*. Milano: Brigola 1879; vol. II, *La filosofia pratica (morale, diritto, politica)*. Milano: Brigola 1883 and vol. III, *La filosofia religiosa, la Critica del Giudizio e le dottrine minori*. Milano: Hoepli 1884. Tocco's main contributions on Kant's philosophy are collected in his volume *Studi Kantiani*, Palermo: Sandron 1909.

4 Felice Tocco, "Del passaggio dalla metafisica della natura alla fisica", in: *Kantstudien* 2, 1898, pp. 66-89. This paper is also available in: *Studi kantiani*, *op. cit*., pp. 213-227.

5 A very useful analysis of this review and a collection of the most important articles published there is available in: Patrizia Guarnieri, *La "Rivista Filosofica" (1899-1908). Conoscenza e valori nel neokantismo italiano*. Firenze: La Nuova Italia 1981.

6 For an overview of Italian philosophy of science during the 20th century see Fabio Minazzi and Luigi Zanzi (Eds.), *La scienza tra filosofia e storia in Italia nel Novecento*. Roma: Istituto Poligrafico dello Stato 1987; Evandro Agazzi (Ed.), *La filosofia della scienza in Italia nel '900*. Milano: Franco Angeli 1986. A critical evaluation of some leading figures is offered by Paolo Parrini, *Filosofia e scienza nell'Italia del Novecen-*

Giovanni Vailati as well as other members of the Peano's school – and last but not least Federigo Enriques – are mentioned in the famous manifesto of the Vienna Circle.[7] Nevertheless, Peano was not a philosopher of science and, broadly speaking, he had no particular interest in philosophy itself, despite his Leibnizian dream of a *characteristica universalis* and despite the enormous influence he exerted on the development of modern logic.[8] By contrast, Vailati was a genuine pioneer and a forerunner of philosophy of science. He was deeply interested in the analysis of scientific language and in history of science. Moreover, his methodological pragmatism assigned a pivotal role to a sophisticated verificationism.[9] But Vailati was at the same time a strong opponent of Kant's theory of knowledge. Like his friend Louis Couturat, he was convinced that the influence of the Kantian philosophy on logic and mathematics had been disastrous and for this reason he praised and emphasized anti-Kantian philosophers such as Bernard Bolzano and Franz Brentano.[10]

Quite different from Vailati, the great mathematician and philosopher of science Federigo Enriques endorsed a Kantian point of view in his attempt to offer a comprehensive exposition of the main *Problems of Science* (this, as is well known, is the title of his great book published in 1906 and promptly translated into French, German, and English).[11] More precisely, Enriques's effort consisted in correcting the evolutionistic positivism of the late 19th century through the Kantian view of knowledge as a constructive process resting on invariant mental structures, although these structures, according to Enriques, were essentially of psychological nature.[12] Interestingly enough, the peculiar understanding of Kant proposed by Enriques resulted in the attempt to conceive of human reason not as determined

to. Figure, correnti, battaglie. Milano: Guerini e Associati 2004.

7 Rudolf Carnap, Hans Hahn, Otto Neurath, *Wissenschaftliche Weltauffassung*. Wien: Artur Wolf Verlag 1929, p. 13. Trans. "The Scientific Conception of the World. The Vienna Circle", in: Otto Neurath, *Empiricism and Sociology* (ed. by Marie Neurath and Robert S. Cohen), Dordrecht: Reidel, 1973, p. 304.

8 Giuseppe Peano, *Opere scelte*. Roma: Edizioni Cremonese, 3 vols., 1957–1959. See also Giulio Giorello (Ed.), *L'immagine della scienza*. Milano: Il Saggiatore 1977.

9 See *Logic and Pragmatism. Selected Essays by Giovanni Vailati*, translated by Claudia Arrighi, ed. by Claudia Arrighi, Paola Cantù, Mauro De Zan and Patrick Suppes, Stanford, CA: CSLI Publications 2010.

10 Giovanni Vailati, "On the Logical Import of the Classification of Mental Facts Proposed by Franz Brentano" (1900), in: *Selected Essays by Giovanni Vailati*, *op. cit*., pp. 107-112. See also Giovanni Vailati, *Scritti*. Firenze: Seeber & Barth 1911, pp. 455, 557, 936.

11 See Federigo Enriques, *Problemi della scienza*. Bologna: Zanichelli 1906, 2nd ed. 1909 (English translation *Problems of Science*, translated by Katharine Royce, with an introductory note by Josiah Royce, London-Chicago: Open Court 1914). Concerning this book and its reception within the philosophy of science in the early 20th century I allow myself to refer to: Massimo Ferrari, *Non solo idealismo. Filosofi e filosofie in Italia tra Ottocento e Novecento*. Firenze: Le Lettere 2006, pp. 205-252.

12 Enriques, *Problemi della scienza*, *op. cit*., pp. 19-20.

by a timeless architecture, but rather as emerging in a continuous development strictly connected with the scientific knowledge. In this sense, Enriques was later on especially engaged by the epistemological analysis of the theory of relativity and of quantum mechanics; his aim was to show how Kant's a priori forms can be interpreted as historical changing forms or, in other words, as relativized a priori components in the scientific endeavour to comprehend reality.[13]

Enriques's dynamics of reason, conceived of in a very broadly Kantian sense, represents a very interesting case study (also regarding his controversial relationship to logical empiricism).[14] Nevertheless Enriques is more an exception than a confirmation of a (supposed) Kantian trend in Italian philosophy of science during the first half of the 20th century. On the other hand, it was only in the period after the Second World War that the philosophy of science experienced a kind of revival and step by step became also an academically recognized discipline. In fact, early in the 1950s the Italian philosophical culture became increasingly sensitive to the philosophy of science that had developed in Europe between the two World Wars and, more recently, became established in the American departments of philosophy as a central feature of modern philosophical inquiry. But the main protagonists of this intellectual adventure as they emerge, for example, from the good survey edited in 1981 by Maria Luisa Dalla Chiara, were not committed to Kant's philosophy.[15] Thus it would be sufficient to recall here the name of Ludovico Geymonat, a former disciple of Peano in Turin and later of Moritz Schlick in Vienna, whose contribution to the rise of philosophy of science in Italy is indisputable, but who in no way can be assimilated to the Kantian tradition.[16]

Nevertheless, as in every interesting story, there exist some remarkable exceptions. The most important one is surely Giulio Preti. He was born in Pavia in 1911 and was a pupil and collaborator of Antonio Banfi, a very original and eminent figure in the Italian philosophy between the Twenties and the Fifties. Banfi was author of a systematic book entitled *Principi di una teoria della ragione* (*Principles of a theory of reason*), in which he endorsed a kind of neo-Kantianism especially influenced by the Marburg School on the one hand and by Husserl's phenomenol-

13 Federigo Enriques, *La théorie de la connaissance scientifique de Kant à nos jours*. Paris: Hermann 1938.

14 See the documentation offered by Michael Stöltzner, "Federigo Enriques e l'Enciclopedia neurathiana", in: *Rivista di storia della filosofia* 54, 1998, pp. 463-494.

15 See Maria Luisa Dalla Chiara (Ed.), *Italian Studies in the Philosophy of Science*. Dordrecht-Boston: Reidel 1981.

16 One of the most important works of Geymonat is the book *Filosofia e filosofia della scienza*. Milano: Feltrinelli 1960, in which Geymonat endorses in agreement with Enriques an historical view of scientific theories. Geymonat's contribution to the diffusion of philosophy of science in Italy is documented in some pioneer works such as *La nuova filosofia della natura in Germania*. Torino: Bocca 1934 and especially *Studi per un nuovo razionalismo*. Torino: Chiantore 1945.

ogy on the other.[17] Banfi labelled his philosophical point of view "critical rationalism", and his main purpose consisted in the elaboration of a critical philosophy of culture. But this attempt was also characterized by a deep interest both in modern science (Banfi wrote important studies about Galileo Galilei) and, to some extent, in contemporary debates on philosophical foundations of mathematical sciences.[18] Although Banfi was well acquainted with eminent neo-Kantian philosophers such as Cassirer or Brunschvicg as well as with the French philosophy of science from Boutroux to Poincaré, he was not, properly speaking, a philosopher of science. To be sure, his transcendental theory of reason was also a theory of scientific reason; and in the 1940s Banfi had a significant role in editing and planning philosophical reviews which gave a remarkable impulse to the establishment of philosophy of science in Italy (so in particular *Studi filosofici*, the very important journal edited by Banfi between 1940 and 1949, and, in part, *Analisi*, a journal conceived by Eugenio Colorni and Ludovico Geymonat before the end of the war).[19] Nevertheless, Banfi represented more of a fruitful philosophical background indebted to the neo-Kantian tradition than an original contribution to the debate on epistemological topics that began to develop in Italy after the Second World War.

3. Giulio Preti: A Kantian turn in philosophy of science

Banfi's legacy was of great importance for his brilliant disciple Giulio Preti. Preti started from the peculiar neo-Kantianism elaborated by Banfi, but already in the early Forties he turned his attention to the Vienna Circle and to logical empiricism, going beyond Banfi's own critical rationalism or, put in other words, transforming it in a new philosophical point of view. (Preti himself later defined this new point of view in an autobiographical sketch, as "historical-objective transcendentalism".)[20] According to Preti, the main failure of Banfi's neo-Kan-

17 See Antonio Banfi, *Principi di una teoria della ragione*. Torino-Milano: Paravia 1926, 3rd ed. Roma: Editori Riuniti 1967.

18 *Ibid*., pp. 99-145. See also Antonio Banfi, "Per un razionalismo critico" (1943), in: Antonio Banfi, *La ricerca della realtà*. Firenze: Sansoni 1959, vol. I, pp. 45-100. Banfi's studies on Galileo are well documented in: Antonio Banfi, *Vita di Galileo Galilei*. Milano-Roma: La cultura 1930, 2nd ed. Milano: Feltrinelli 1962 and Antonio Banfi, *Galileo Galilei*. Milano: Ambrosiana 1949, 2nd ed. Milano: Il Saggiatore 1961.

19 On Banfi's philosophical activity as editor of *Studi filosofici* see Eugenio Garin, *Intellettuali italiani del XX secolo*. Roma: Editori Riuniti 1974, pp. 241-264. To keep in mind is also Banfi's short but important article "Appunti per una metodologia critica", in: *Analisi* 1, 1945, pp. 25-35 (also published in: Banfi*, La ricerca della realtà*, *op. cit*., vol. I, pp. 191-201).

20 Giulio Preti, "Il mio punto di vista empiristico" (1958), in: Giulio Preti, *Saggi filosofici*, edited by Mario Dal Pra, Firenze: La Nuova Italia 1976, vol. I, p. 486. On Preti's philosophical development and his philosophy see Fabio Minazzi, *Giulio Preti: bibliografia*. Milano: Franco Angeli 1984; Mario Dal Pra, *Studi sull'empirismo logico*

tiansm (and also of Cassirer's) consisted in missing the role of experience both as an autonomous source of knowledge and as an unavoidable point of reference in order to ascertain whether our linguistic propositions are meaningful or not.[21] The logical empiricist criterion of verification, particularly in its liberalized version and enlarged through Dewey's pragmatism, became in this way the new, central aspect of Preti's philosophy of science in the early Fifties.[22] Dewey, in particular, was in Preti's view the paradigmatic philosopher of a new scientific culture and, at the same time, the author of a philosophical perspective conceiving of verification as "the process by which a statement is constructed as a true statement".[23] On the other hand, Preti was convinced that logical empiricism and pragmatism failed in missing the crucial role played by conceptual schemes, categorical structures or, more precisely, linguistic forms in the constitution of experience, a role that made it possible for science to be really "rigorous".[24]

The neo-Kantian view of reason was in this sense still valid, but it required a re-formulation that Preti sought to articulate in two different dimensions. First and foremost, Preti attempted to transform the a priori forms of reason in the linguistic dimension. In this he was especially influenced both by Carnap's philosophy of language and from Morris's conception of an unavoidable enlargement from semantic to pragmatic of language.[25] Still before Karl-Otto Apel coined his famous motto, Preti was therefore able to sketch a "semiotic transformation of Kantianism", according to which "the problem of knowledge [in the old Kantian sense] is a semantic one" – precisely because the structures of knowledge and its categorical frameworks are of linguistic nature: knowledge itself is primarily manifested in "speech (*discorso*)".[26] Second, Preti emphasized even more than Banfi or Cassirer the historical, always changing nature of the a priori forms of reason. Thus he spoke explicitly of a "historical, relativized a priori", which ena-

---

di Giulio Preti. Napoli: Bibliopolis 1988; P.L. Lecis, *Filosofia, scienza, valori: il trascendentalismo critico di Giulio Preti*. Napoli: Morano 1989; Fabio Minazzi (Ed.), *Il pensiero di Giulio Preti nella cultura filosofica del Novecento*. Milano: Franco Angeli 1990; Paolo Parrini and Luca M. Scarantino, (Eds.), *Il pensiero filosofico di Giulio Preti*. Milano: Guerini e Associati 2004; Luca M. Scarantino, *Giulio Preti. La costruzione della filosofia come scienza sociale*. Milano: Bruno Mondadori 2007.

21 See for example Preti's programmatic article "Il razionalismo contemporaneo", in: *Inventario* 3, 1949, pp. 108-117.

22 Preti, "Le tre fasi dell'empirismo logico" (1954), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. I, p. 296.

23 Preti, "Dewey e la filosofia della scienza" (1951), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. I, p. 86.

24 *Ibid*., p. 96.

25 Preti, "Linguaggio comune e linguaggi scientifici" (1953), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. I, pp. 208-209.

26 Preti, "Il linguaggio della filosofia" (1962), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. I, p. 462.

bles the construction and delimitation of different "regional ontologies".[27] This was for Preti the core of a new epistemology that emerged from two quite different traditions, neo-Kantianism and logical empiricism. As Preti claimed, epistemology as reflection of the science about itself is always *historical* epistemology.[28] This means not only that Preti conceived epistemology as historical oriented, but that the *historical dimension* is, from his standpoint of view, a proper and essential feature of epistemology as such. Science and scientific knowledge are always subject to historical development, change, and even revolution. It is exactly for this reason hat epistemology too is deeply rooted in history and it can in no way aim to circumscribe – as Kant did – universal and necessary forms of knowledge.[29]

Preti, we thus can say, anticipated to a considerable extent a perspective which is today quite familiar to us both from epistemological and from historical points of view – a "dynamics of reason" at the crossroad between logical empiricism and neo-Kantianism.[30] Not accidentally, it is a student of Preti – Paolo Parrini – who has offered in the last decades important contributions to a new interpretation of logical empiricism from a Kantian or neo-Kantian standpoint, thereby renewing and modifying at the same time Preti's own perspective.[31]

4. Preti on historical knowledge

A very interesting aspect to stress here is that Preti, in agreement with his general epistemological program, drew attention also to neglected topics such as the problem of historical knowledge – an issue which Preti dealt with extensively in *Praxis ed empirismo*, his highly original book published in 1957.[32] In Preti's opinion the traditional opposition, that is logical empiricism *versus* historicism and historical knowledge, was only the outcome of a misleading conception of history in general. History is a regional ontology, or more precisely both a formal and a material

27 Preti, "L'ontologia della regione «natura» nella fisica newtoniana" (1957), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. I, p. 416.

28 Preti, "Due orientamenti nell'epistemologia" (1950), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. I, pp. 54, 76.

29 Preti, "Il mio punto di vista empiristico", *op. cit*., vol. I, pp. 493-495. See also Preti, *Morale e metamorale. Saggi filosofici inediti*, edited by E. Migliorini, Milano: Franco Angeli 1989, p. 91.

30 We refer obviously to Michael Friedman, *Dynamics of Reason. The 1999 Kant Lectures at Stanford University*. Stanford, CA: CSLI Publications 2000.

31 See Parrini, *Knowledge and Reality. An Essay in Positive Philosophy*. Dordrecht: Kluwer 1998 and *L'empirismo logico. Aspetti storici e prospettive teoriche*. Roma: Carocci, 2002.

32 See Preti, *Praxis ed empirismo*. Torino: Einaudi 1957; new edition with a Preface by Salvatore Veca and an Afterword by Fabio Minazzi, Milano: Bruno Mondadori 2007 (I quote from this last edition).

ontology[33], whose categorical frameworks, or meanings in Dewey's sense, have to be established through the human mind, according to the more general procedures describing the conditions of possibility of all the sciences, the *Naturwissenschaften* as well as the *Geisteswissenschaften*.

In ways quite different from Popper, who he criticized briefly but incisively, for Preti it was possible to analyse and to construct the language of historical inquiry as a scientific language.[34] This means, first, that historical knowledge too adheres to the principle of verification: the knowledge of the past is not radically different from scientific knowledge in general. Every fact we are dealing with is a fact inferred at the present moment in that we verify that a similar fact – so, e.g., a trace of the past – really exists.[35] Secondly, this means, and at the same time presupposes, that it is possible to formulate historical causal statements, that is statements maintaining the lawfulness of a specific chain of events, without thereby assuming a metaphysical plan in the development of history.[36] Finally we can formulate predictions concerning historical events, inasmuch as the inference from the present to the future is quite similar to the inference from the present to the past, i.e. it is a "probable induction" whose verification depends on the "data" of present".[37] Thus historical statements belong, according to Preti, to the family of statements that Dewey classified as possessing "warranted assertibility". It was in this sense that Preti considered historical knowledge as a part of human knowledge, aiming to explain "human situations" through causal statements.[38]

On the other hand, Preti was particularly interested in giving an account of historical knowledge as a peculiar framework of categories or meanings circumscribing the "regional ontology" that is *history*. History, says Preti, has its specific "meanings" too, meanings which enable us to recognize a fact as an "historical fact".[39] What is even more stimulating in this account is that Preti conceived of similar meanings or conceptual frameworks as historically changeable forms, that is as general "paradigms" which make it possible that historical knowledge as scientific knowledge deals with the boundless quantity of facts or events belonging to the human world.[40] These facts or events need to be organized and selected according to a "paradigm": for instance, historical materialism endorses certain

33 *Ibid*., p. 124.
34 *Ibid*., p. 123.
35 *Ibid*., pp. 125-126.
36 *Ibid*., pp. 126-127.
37 *Ibid*., p. 127.
38 *Ibid*., pp. 128-129.
39 *Ibid*., p. 130. It is noteworthy that Preti approaches the problem of historical knowledge in terms that are quite near to the typical neo-Kantian solution, in particular in the sense of Heinrich Rickert and, to some extent, of Max Weber. Preti was well acquainted with this perspective, nevertheless he quotes in this context neither Rickert nor Weber.
40 *Ibid*., p. 134.

principles of organisation of human facts in order to systematize historical knowledge in the same way in which Newton's principles of dynamics are capable to systematize classical physics.[41] Preti was therefore particularly sensitive in giving an account of the structural affinities between historical knowledge and scientific knowledge. In particular, he stressed that historical knowledge too needs linguistic conventions. Values, practical purposes, universes of discourse regarding art, religion, science, and culture in general are the outcome of a long historical experience and development making it possible that we speak of rupture or even of revolution in human history. Now, what is important or crucial in history depends on the convention we adopt in order to distinguish new and old, continuity and discontinuity, normal and revolutionary. But this last aspect has to do with a more comprehensive view of historical development, or – put in other words – with the idea that history is characterized by the tension between continuity and discontinuity, between tradition and innovation or *revolution*.

History, Preti maintains, means on the one hand the emergence of novelty, "the creative moment" of human activity; but on the other hand history would be simply impossible without the "connective tissue" of tradition.[42] Moreover, Preti attempted to apply a similar conception of history within the history of culture and, more precisely, within the history of thought. So in his essay on Newton, in which he presented the "regional ontology" of nature according to the principles of the Newtonian physics, Preti considered the scientific revolution accomplished by Newton as a genuine revolution (in a sense quite close to Alexandre Koyré's interpretation of Newtonian science) and he tried to show that the relativistic and post-Newtonian physics did not mean a revolutionary change similar to the Newtonian one. Preti's main concern lay in drawing attention to the highly meaningful historical fact that "every change in our knowledge happens in a traditional framework, in a traditional universe of discourse, in a traditional language, with respect to which this change introduces novelties that not only do not distort, but rather presuppose that material ontology already given."[43] And he added: "This ontology of nature is the result of the structure itself of the scientific language, of the methodological postulates, of the rules both of discourse and verification which the scientific tradition of the West has selected and elaborated."[44]

41 *Ibid*., p. 138.

42 *Ibid*., pp. 147-148.

43 Preti, "L'ontologia della regione «natura» nella fisica newtoniana" (1957), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. I, p. 420.

44 *Ibid*., p. 434.

5. Preti on the history of philosophy

This was surely an original insight not only within Italian philosophy of science. The historical turn we usually associate with the post-neopositivistic philosophy of science was accomplished by Preti already in the 1950s and was, to some extent, a consequence of his acquaintance both with the Italian tradition of historicism in general (although Preti rejected the metaphysical implications of this tradition) and with the history of philosophy in particular. Moreover, it is highly illuminating to consider that Preti, some years before Thomas Kuhn, devoted his attention to the historical dynamics of scientific thought and, in particular, to the role played by "parameters" in order to justify continuity and discontinuity within the history both of science and philosophy. In his essay "Continuità e discontinuità nella storia della filosofia (*Continuity and Discontinuity in History of Philosophy*)" published in 1951, Preti maintained that philosophy, as theoretical enterprise, exists only "in historical form".[45] But this statement signifies in no way that philosophy is a sort of continuous development of an uninterrupted tradition excluding novelties and breakdowns. On the contrary, Preti stressed very clearly how philosophy can be characterized as an historical process resting on *parameters*, that is on categories which are historical changeable. And it is exactly this change of parameters in the history of philosophy that represents the discontinuity breaking its apparently intangible continuity.[46] "In this change", Preti asserted, "consists what we call a *philosophical revolution*, the opening of a new age after the crisis of the previous age. Nevertheless, a similar change begins always with a critique of the previous forms of thought, with a new statement of the same problems (a new statement which only later is discovered as a veritable position of new problems): it is exactly through this kind of 'chain' that we can see within the history of philosophy a continuity allowing us to keep at least the unity of the name 'philosophy."[47]

Preti was thus fully aware of the fact that the history of philosophy (and history of science too) can be interpreted as a history of philosophical (and scientific) "traditions".[48] Put in other words, "we are living within a tradition."[49] Within traditions discontinuities or even revolutions are constitutive aspects of their historical development. Here Preti attempted to show how philosophical (or scientific) parameters change. More precisely, in the history of philosophy there can be detected "essences" (realism, idealism, rationalism, empiricism and so on), which are useful in order to organize and to unify philosophical trends and positions

45 Preti, "Continuità e discontinuità nella storia della filosofia" (1951), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. II, p. 233.

46 *Ibid*., p. 237.

47 *Ibid*., pp. 237-238.

48 Preti, "Continuità ed «essenze» nella storia della filosofia" (1956), in: Giulio Preti, *Saggi filosofici*, *op. cit*., vol. II, p. 249.

49 Preti, "Il mio punto di vista empiristico", *op. cit*., p. 484.

from an historical point of view.[50] The movement from one "essence" to another signifies properly a change or even a revolution in history of philosophy. At the same time, to be a rationalist or an empiricist, a materialist or a spiritualist means to share a "logic" or a certain way of thought – but we can simply say, to share a "parameter" or, in Kuhn's language, a "paradigm".[51]

It was no accident therefore that Preti was very sensitive to Otto Neurath's famous metaphor of the sailors travelling in a boat in open sea and thus compelled to repair the boat without any possibility to come back to the harbour. According to Preti, Neurath's metaphor for the impossibility of a *tabula rasa* was valid not only in the case of scientific knowledge, but also for philosophy as such, given the impossibility to promote a philosophy (and in particular a philosophy of culture) destroying or ignoring the tradition within which it is grown.[52]

6. Conclusion

Preti's historical view of the double movement of continuity and discontinuity characterizing the evolution of philosophical traditions can be labelled, broadly speaking, as a sort of pre-Kuhnian way of considering the dynamics of philosophical thought.[53] Preti's own position regarding this last aspect can be therefore appreciated and reappraised as an original contribution to the elucidation of the "structure of philosophical revolutions". This testifies to the fact that Preti was a very creative philosopher, certainly the most important figure within the Italian philosophy of science in the second half of 20th century.[54] Unfortunately, Preti was not an English-speaking philosopher and his work, with the only exception of a French translation of some of his papers (prompted by Jean Petitot), is largely ignored outside the Italian philosophical culture.[55] It would be surely a great homage

50 Preti, "Continuità ed «essenze» nella storia della filosofia", *op. cit*., p. 248.

51 *Ibid*., p. 250.

52 Preti, *Lezioni di filosofia della scienza (1965-1966)*, edited by Fabio Minazzi, Milano: Franco Angeli 1989, p. 153 n. 16.

53 See also Scarantino, *Giulio Preti. La costruzione della filosofia come scienza sociale*, *op. cit*., p. 303.

54 See Parrini, *Filosofia e scienza nell'Italia del Novecento*, *op. cit*., p. 199.

55 See Preti, *Écrits philosophiques. Les Lumières du rationalisme italien*, textes choisis et presentés par Luca M. Scarantino, préface per Jean Petitot, Paris: Cerf 2002. Contributions on Preti's philosophy have been also published in: *Diogène*, n. 216, octobre-décembre 2006. In his recent talk "Is there a European Philosophy of Science?" delivered at the International Conference "Philosophy of Science in Europe – European Philosophy of Science and the Viennese Heritage" (Vienna, Dec. 7, 2011), Gereon Wolters has drawn the attention to the work of "the very interesting Italian philosopher of science Giulio Preti (1911-1972), to whom we owe among other things a fascinating pragmatist and cultural embedding of philosophy of science". Wolters rightly apologizes that "unfortunately, not a line seems to have been translated into

to this highly innovative figure of contemporary European philosophy of science to promote a deeper acquaintance with his very illuminating papers and books. And while, sociologically speaking, Preti may be still considered as a marginal figure in 20th century European philosophy of science, it is noteworthy to recall that marginal in that sense in no way means either second-class or belonging exclusively to the past.

Department of Philosophy
University of Turin
Via S. Ottavio 20
10124, Torino
Italy
massimo.ferrari@unito.it

English" (p. 11: quoted from the manuscript by kind permission of the author). We are in fully agreement with Wolters's *lamentatio* against the *globalized parochialism* of Anglophone philosophy of science.

Thomas Mormann

Topology as an Issue for History of Philosophy of Science

Abstract

Since antiquity well into the beginnings of the 20th century geometry was a central topic for philosophy. In contrast, most philosophers of science, if they took notice of topology at all, considered it as an abstruse subdiscipline of mathematics lacking philosophical interest. Here it is argued that this neglect of topology may be conceived of as the sign of a conceptual sea-change in philosophy of science that expelled geometry, and, more generally, mathematics, from its central position in philosophy of science and, instead, placed logic at center stage in the 20th century philosophy of science. Only in recent decades logic has begun to loose its monopoly and geometry and topology received a new chance to find a place in philosophy of science, as an object for philosophical reflection and as a conceptual tool for doing philosophy.

1. Introduction

From antiquity to the beginnings of the 20th century philosophers took geometry as the paradigmatic example of science. Geometry defined what was to be considered as scientific knowledge. "More geometrico" was considered as a sign of quality for philosophical and scientific argumentation. At the beginning of the 20th century, the privileged philosophical status of geometry seemed to be as solid as it always had been. For philosophers such as Russell, Cassirer or Carnap, to name but a few, philosophical problems posed by geometry played a central role in their investigations – at least at the beginnings of their careers:

(i) Russell started his philosophical career in 1897 with the dissertation *The Foundations of Geometry.*[1] A few years later, in *The Principles of Mathematics* he treated themes from geometry at great length.[2] In *The Analysis of Matter* Russell was engaged in using topological methods for the "logical analysis" of space and time.[3]

1 Bertrand Russell, *The Foundations of Geometry*. Cambridge: Cambridge University Press 1897.

2 Russell, *The Principles of Mathematics*. Cambridge: Cambridge University Press 1903.

3 Russell, *The Analysis of Matter*. London: Routledge 1927.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_34,

(ii) Throughout his life, Cassirer considered *Klein's Erlangen Programme* as a guideline for the epistemology of his "Critical Idealism" characterizing the task of epistemology as finding the ultimate invariants of scientific knowledge. In *Substanzbegriff und Funktionsbegriff* and much later in *The Philosophy of Symbolic Forms* he dedicated central chapters to concept formation in geometry which he considered as a paradigmatic case for concept formation in science *überhaupt*.[4]

(iii) Carnap's first philosophical publication was his dissertation *Der Raum. Ein Beitrag zur Wissenschaftslehre*.[5] There he sought to establish the topological structure of space as a modernized version of a Kantian *synthetic a priori*. Moreover, the geometrical considerations of this work may be regarded as an important source for his later philosophy.[6]

The high esteem of 20th century philosophy of science for geometry and, more generally, for mathematics, went well beyond the philosophical currents that in the following decades were to form analytic philosophy of science. For instance, also in phenomenology great emphasis was put on geometry as a paradigmatic example of scientific knowledge. This is evidenced not only by the work of Husserl himself but also by the contributions to a phenomenological philosophy of mathematics by mathematicians and philosophers such as Hermann Weyl, Dietrich Mahnke or Oskar Becker.[7] The same holds for some currents of Neokantian philosophy of science, for instance, the Marburg school of Neokantianism whose members, such as Hermann Cohen, Ernst Cassirer, and Paul Natorp, emphasized the role of mathematics in many works.

In sum, in the early decades of the last century, geometry certainly did *not* belong to the sciences "neglected by received philosophy of science" – on the contrary, at that time geometry was one of the hot topics of received philosophy of science.

---

4 Ernst Cassirer, *Substanzbegriff und Funktionsbegriff: Untersuchungen über die Grundfragen der Erkenntniskritik*. Berlin: Bruno Cassirer 1910, trans. *Substance and Function*. Chicago and LaSalle: Open Court 1923; *Philosophie der symbolischen Formen*, 3 vols., Berlin: Bruno Cassirer 1923–29, trans. *The Philosophy of Symbolic Forms*. New Haven: Yale University Press 1953.

5 Rudolf Carnap, *Der Raum. Ein Beitrag zur Wissenschaftslehre Kantstudien Ergänzunghefte 56*, 1922.

6 See Thomas Mormann, "Geometrical Leitmotifs in Carnap's Early Philosophy", in: Michael Friedman, Richard Creath (Eds.), *The Cambridge Companion to Carnap*. Cambridge: Cambridge University Press 2007, pp. 43-64.

7 Hermann Weyl, *Das Kontinuum*, 1919, trans. *The Continuum*. New York: Dover 1994. Dietrich Mahnke, "From Hilbert to Husserl: First Introduction to Phenomenology, Especially that of Formal Mathematics". Translated by D. Boyer. *Studies in the History and Philosophy of Science* 8, 1977, pp. 71-84 (orig. 1923). Oskar Becker, "Beiträge zur phänomenologischen Begründung der Geometrie und ihrer physikalischen Anwendungen", in: *Jahrbuch für Philosophie und phänomenologische Forschung* 4, 1923, pp. 385-560.

This was soon to change, however. While geometry as a methematical discipline experienced a golden age during the 20th century mathematics, it lost its privileged status in philosophy. This became apparent first by the fact that traditional geometry's most promising offspring – topology – fell into philosophical disregard. The philosopher's traditionally high appreciation of geometry did not extend to topology as its modern successor. On the contrary, in the 20th century topology may be rightly characterized as a science "neglected by received philosophy of science". Even more, the philosophical neglect of topology was just the harbinger of a fundamental sea-change in philosophy of science, namely, the substitution of geometry, and more generally of mathematics, as a core issue of philosophy of science, by logic. Painted with a broad brush the 20th century mainstream logical empiricist 20th century philosophy of science was a *logic-centered* philosophy of science, concentrating on logical questions concerning the logical structure of science.[8]

Since from the mathematical point of view there is no essential epistemological, ontological, or methodological difference between geometry and topology, the negligible amount of attention that philosophy paid to topology in the last century must be attributed to a change in the way philosophers understood the aims and methods of philosophy of science. This renders philosophy's neglect of topology an intricate problem for the *history of philosophy of science*.

For the following it is useful to distinguish between two different aspects according to which the relations between traditional philosophy and geometry on the one hand, and between 20th century philosophy and topology on the other, differed from each other:

First, 20th century philosophy of science showed no interest in topology as an object of philosophical reflection. There has been no "philosophy of topology" in analogy to disciplines such as "philosophy of physics", "philosophy of biology", or "philosophy of geometry" (as it existed as a living philosophical discipline till the beginning of the last century). Second, traditionally geometry had also served as a source for inspiration and as an arsenal of conceptual tools for philosophy itself. This fruitful exchange did not find a continuation between the 20th century philosophy of science and topology. Ideas from topology hardly found their way in the conceptual tool kit of the philosopher of science.

8 During the decades the concentration of philosophy of science on the logical aspects of science was assessed quite differently: in the 1930s we find Carnap's sweeping claim that "philosophy of science just is logic of science". At the end of the 20th century Carnap's thesis had lost some of its appeal – to put it mildly. Van Fraassen put forward the harsh verdict: "It was a tragedy for philosophers of science to go off on these logico-linguistic tangles, which contributed nothing to the understanding of either science or logic or language." (Bas C. van Fraassen, *Laws and Symmetry*. Oxford: Clarendon Press 1989, p. 221). This is not to say that the use of mathematics provides a foolproof method for doing substantial philosophy of science.

My thesis is that this twofold neglect of topology by philosophy of science was just the first sign of a fundamental sea-change in philosophy of science, namely, the replacement of mathematics as a guiding science for philosophy by logic. Although the core disciplines of science were mathematized sciences, mainstream philosophy of science was to treat science from an exclusively logical point of view. The disregard of mathematical, in particular geometrical and topological, aspects of science by philosophers of science was in stark contrast to the emphasis that they put on the logical aspects of the scientific enterprise.

2. Topology as a problem for philosophy of science

What is topology? It goes without saying that a short paper like this is not the appropriate place for answering this question.[9] As a mathematical discipline in its own right, recognizable also for non-mathematicians, topology came into being around the turn of the last century. Let us mention the names of Cantor, Poincaré, Frechet, and Hausdorff, to name just a few of the leading figures. Topological ideas and problems may be traced back, however, to Leibniz and Euler: one may think of the famous "Seven bridges of Königsberg" or "Euler's theorem" dealing with the relation between the vertices, edges, and faces of polygons, from which Lakatos squeezed so much juice for philosophy of mathematics.[10]

In broadest outline, then, topology is concerned with the conceptual analysis of spatial notions, such as "space in general", "connectedness", "neighborhood", "approximation", "convergence", "continuity", "mappings", "transformations", "boundedness", and many others. Evidently, these concepts may have had their origin in our daily experiences with physical space but they make sense far beyond the original Euclidean frame.

A first step to overcome the traditional Euclidean conception of space was to consider general metrical spaces:

(2.1) Definition. A metrical space (X, d) is a set of points endowed with a distance function d: X x X → **R (R** the real numbers) satisfying the axioms:

---

9 The reader may find some preliminary answers of this question especially adapted to the needs of philosophers in papers by Franklin and by Grosholz (Philip Franklin, "What is Topology?", in: *Philosophy of Science* 2, 1935, pp. 39-47; Emily Grosholz, "Two Episodes in the Unification of Logic and Topology", in: *The British Journal for the Philosophy of Science* 36, 1985, pp. 147-157.), or in the monographic issue of *The Monist*, "Topology for Philosophers" (Barry Smith and Wojciech Zelaniec (Eds.), *Topology for Philosophers*, *The Monist* 79, 1, 1996). Stephen Willard, *General Topology*. New York: Dover 2004 (orig. 1970) offers a classical introduction for mathematically interested readers.

10 Imre Lakatos, *Proofs and Refutations*. Cambridge: Cambridge University Press 1976.

(i) $d(x, x) = 0$.
(ii) $x \neq y \Rightarrow d(x, y) = d(y, x) > 0$.
(iii) $d(x, y) + d(y, z) \geq d(x, z)$ (Triangle Inequality). ♦

The requirements (2.1) (i) and (ii) may be considered as almost analytical for any reasonable notion of distance, while (2.1) (iii) rather faithfully reflects a property of the Euclidean distance function.

A further, more radical step away from traditional geometry toward topology in its proper sense, was the generalization from metrical to general topological spaces that freed the topological, i.e. the spatial, from any vestige of a quantitative metric or distance function.

In the literature a variety of equivalent definitions of a topological space exists. Arguably, the following is the most common one:

(2.2) Definition. Let X be a set and denote by PX the power set of all subsets of X. A topological space (X, OX) is defined as a set X with a class $OX \subseteq PX$, called the open sets of the topological space, that satisfy the following requirements:

(i) X and the empty subset Ø are open sets.
(ii) The union of any collection of open sets is open.
(iii) The intersection of two open sets is open.

OX is called a topology or a topological structure on X. ♦

A metrical space (X, d) such as the Euclidean space is rendered a topological space by defining the metrical topology as the one that is generated by the "open balls":

$$U(x, \varepsilon) := \{y; d(x, y) < \varepsilon, x \in X \text{ and } \varepsilon > 0\}.$$

It should be noted that in general a set X, in particular the set of points of Euclidean space, can be endowed with many different topological structures OX. Among the many possible topologies on a set X one may mention the coarsest topology defined as $O_0X = \{Ø, X\}$ and the discrete topology defined as $O_1X = PX$. All other topologies OX on X are "between" these two extreme topologies. More precisely, the topologies OX on X can be partially ordered by set-theoretical inclusion:

$$\{Ø, X\} = O_0X \subseteq OX \subseteq O_1X = PX$$

It would be an egregious error to take the profusion of possible topologies OX on X as evidence that the concept of topology is arbitrary and therefore trivial. The point of defining a topological structure on a set X is not to define just any one, but rather to define an interesting one. What is to be considered as an interesting topological structure highly depends on the specifics of the situation. It requires considerable skill and mathematical ingenuity to find "good" topologizations and to exploit them in a fruitful manner.

One of the early masters of this "art of topologizing" was the American mathematician Marshall H. Stone who obtained spectacular results by applying the new topological devices in many areas of mathematics, in particular in lattice theory and functional analysis.[11] Stone coined the maxim "*You must always topologize*".[12] He conceived of topology as a universal *method* or *perspective* from which every mathematical problem should be looked at, i.e. all objects should be considered as topological ones. The topological was a kind of a general *a priori* form, under which mathematical objects and relations were to be perceived in order to reveal their essential aspects.

Perhaps one may say that Stone sought to conceive of topology as a generalized "transcendental aesthetics" roughly in Kant's sense, based on a general topological *a priori*. The fruitfulness of Stone's topological was amply demonstrated in many areas of 20th century mathematics.[13] Nevertheless, among philosophers his work has remained virtually unknown up to this day.

Since the axioms for a topological structure are extremely general, it is not to be expected that from them strong specific results can be obtained. Rather, an important task of topology is to single out appropriate special classes of topological spaces for which one can prove more specific results. For instance, the already mentioned metrical spaces are an important class, metrizable spaces and Hausdorff spaces provide more general classes, among many others.[14]

Studying a topological space in isolation seldom yields interesting results. Rather, relations between topological spaces are of crucial importance. Hence the second fundamental concept of topology, which has to be mentioned, is that of a continuous map between spaces:

(2.3) Definition. Given two topological spaces X and Y a set-theoretical map $X—f→Y$ is called continuous (with respect to the topologies OX and OY defined on X and Y, respectively), if and only if for every $B \in OY$ the inverse image $f^{-1}(B) = \{a; f(a) \in B\}$ is an element of OX. Roughly, then, (set-theoretical) topology may be described as the theory of topological spaces and continuous maps between topological spaces.

After these preparations some important general types of topological problems may be described as follows:

---

11 Marshall H. Stone, "The Theory of Representations for Boolean Algebras", in: *Transactions of the American Mathematical Society* 44, 1936, pp. 807-816.

12 Mario Piazza, " 'One Must Always Topologize' ": Il teorema di Stone, la 'topologia influente' e l'epistemologia matematica", in: *Rivista di storia della scienza* (ser. II) 4, 1995, pp. 1-24.

13 Peter Johnstone, *Stone Spaces*. Cambridge: Cambridge University Press 1982.

14 For a comprehensive classification of types of topological spaces and their logical relations the reader may consult the very useful compilation in Lynn Arthur Steen, J. Arthur Seebach Jr., *Counterexamples in Topology*. New York: Springer 1978.

(i) Given topological spaces X and Y, can one prove (or disprove) that they are "equivalent" in a sense to be specified?
(ii) Given topological spaces X and Y, does there exist a non-trivial continuous map X—f→Y?
(iii) Can one find interesting invariants that can be calculated to characterize topological spaces in an efficient way (e.g. fundamental groups, higher homotopy groups, (co-) homology theories)?

For many, apparently "elementary" spaces these problems are still unsolved today.

3. Philosophers and topology: Some examples

Let us now briefly mention some of the few philosophical attempts to come to terms with topology. The most important example is certainly Russell but he was not the only philosopher who was interested in topology. For instance, Carnap in his dissertation *Der Raum* had proposed to save a Kantian *synthetic a priori* of space by conceiving of the metrical structure of space as a mere convention but retaining the topological structure of Euclidean space as a core *a priori*. This proposal seems not to have impressed his fellow philosophers. Moreover, Carnap himself gave it up soon after the publication of *Der Raum*. In the *Aufbau* traces of topology are still noticeable, but in his later work in philosophy of science topology and geometry does no play a role at all.[15] Cassirer emphasized in his philosophy of science the importance of geometry for philosophy of science, but offered only some general, passing remarks on the role of topology.[16] Compared with Carnap's and Cassirer's remarks Russell's topological project was by far as the most sustained and detailed one. Russell developed his topological ideas with various degrees of precision and explicitness in several contributions, beginning with *Our Knowledge of the External World*, later in a more detailed way in *The Analysis of Matter*, and finally in *On Order in Time*.[17] Indeed, Russell sought to use the methods of topology for the core task of scientific philosophy, to wit, for logical analysis.

According to Russell the aim of logical analysis was the elimination of suspicious or otherwise undesired entities from philosophical discourse. In *Our Knowledge of the External World* he sought to show by means of examples,

15 See Mormann, *op. cit*.
16 See Cassirer, *Philosophy of Symbolic Forms*, *op. cit*., pp. 422-423.
17 Russell, *Our Knowledge of the External Worlds as a Field for Scientific Method in Philosophy*. London: Routledge and Kegan Paul 1914; *The Analysis of Matter*, *op. cit*.; "On Order of Time", in: Russell, *Logic and Knowledge*. London: Routledge 1956, pp. 347-363 (orig. 1936).

the nature, capacity, and limitations of the logical-analytic method in philosophy. … The central problem by which I have sought to illustrate method is the problem of the relation between the crude data of sense and the space, time and matter of mathematical physics.[18]

In other words, Russell proposed to apply topology as a means for the solution of a genuine philosophical problem, namely, the logical analysis and the elucidation of the complex relation between sense data and the mathematical conceptualizations of physics.

More precisely, Russell wanted to show that the basic mathematical structures of physical space-time – usually conceived of as structured sets of spatial and temporal points (instants) – could be logically reconstructed from 'crude sense data', later to be characterized as 'events'. He credited Whitehead with the basic ideas of this approach:

I owe to Dr. Whitehead the definition of points, the suggestion for the treatment of instants and "things", and the whole conception of the world of physics as a *construction* rather than an *inference*. What is said on these topics here is, in fact, a rough preliminary account of the more precise results which he is giving in the fourth volume of our *Principia Mathematica*.[19]

Regrettably, the announced fourth volume of *Principia Mathematica* never saw the light of the day. In *Process and Reality* Whitehead put forward something like a topological philosophy, but it was not more than a sketch and had no influence on mainstream analytic philosophy of science.

A more detailed account of the construction of points can be found in *The Analysis of Matter*; Russell's last original work on the matters of points (more precisely on temporal points, i.e. instants) was "On Order in Time". For instants as well as for spatial points Russell used the same constructional method. His paper opens with the following contention:

[I]nstants are mathematical constructions, not physical entities. If, therefore, there are instants, they must be classes of events having certain properties. For reasons explained in *Our Knowledge of the External World*, pages 116-120, an instant is most naturally defined as a group of events having the following two properties:
(1) Any two members of the group overlap in time, i.e. neither is wholly before the other.
(2) No event outside the group overlaps with all of them.[20]

Intuitively, Russell's sketchy construction of an instant may be described as an "onion construction", i.e., Russell defines an instant as the limit of a nested collection of temporal intervals. In modern terms, Russell's construction resembles a construction of instants by maximal filters. In mathematically rigorous terms, such

18 Russell, *Our Knowledge of the External World*, *op. cit*., p. 10.
19 *Ibid*., pp. 10-11.
20 Russell, "Order in Time", *op. cit*., p. 347.

constructions were carried out by Tarski and Stone around the same time. The crucial point in this construction was the existence of maximal filters (ultra-filters). Their existence can only be ensured by the axiom of choice or a similar principle as also Russell had noted. In other aspects, however, Russell's constructions remained vague and even seem to be mathematically flawed in some respects.[21]

Evidence that philosophy of science had actually lost contact with topology was that neither Russell nor any other philosophers of science ever took notice of the path-breaking work of the American mathematician Marshall H. Stone who in the 1930s proved one the most important theorems of the 20th century mathematics, to wit, Stone's representation theorem. This theorem established surprising and deep relations between logic and topology, and it could have easily been used to secure what Russell sought to achieve, namely the construction of (temporal and spatial) points from temporal intervals and spatial points from spatial regions.

Despite his prominence as a philosopher, Russell's excursions into topology did not arouse much interest among his colleagues. In the following decades philosophers now and then paid due reference to the later Russell's talent for dealing with the conceptual tools of topology but his project did not find followers. Worse, no philosopher realized that Russell's sketch of a topological logical analysis had long been superseded by the ongoing evolution of topology. It never occurred to the mainstream philosophers of science that meanwhile mathematicians had produced much better topological tools than those that Russell had vaguely adumbrated. Topological and geometrical methods in philosophy of science showed up again only much later, and in a context quite unrelated to Russell's original project.

4. The return of mathematical methods in philosophy of science

Russell's attempt to introduce topological methods in philosophy of science for the logical analysis of philosophical and scientific notions remained unsuccessful. Under the reign of a reductionist logical philosophy mathematics, and *a fortiori*, geometry and topology, was doomed to be considered as irrelevant as an object and as a tool of philosophy of science:

(i) Being allegedly reducible in one way or other to logic, mathematics ceased to be an interesting object for the investigations of philosophy of science in its own right. Rather, from the perspective of logical philosophy of science, the only philosophically interesting area of mathematics was its logical foundation. The huge rest of "real mathematics" was considered as philosophically uninteresting,

21 See Mormann, "Russell's Many Points", in: A. Hieke and H. Leitgeb (Eds.), *Reduction, Abstraction, Analysis*, Proceedings of the 31th International Wittgenstein Symposium in Kirchberg 2008, pp. 239-258, 2009.

being relevant only for mathematicians or those who were concerned with its applications of mathematics.

(ii) From the perspective of a strictly logical philosophy of science any project to use mathematics as a *tool* for philosophy of science made no sense. Tapping the conceptual sources of geometry and topology for understanding and elucidating the structure and the function of empirical theories seemed pointless. Rather, the only legitimate tool for philosophy of science was logic.

Through the decades the conceptual limitations of a strictly logical approach to philosophy of science became more and more visible. Even an arch logical empiricist such as Carnap came to admit that that logic of science might not be everything that philosophy of science had to say about science. He therefore proposed a kind of division of labour that proposed to complement the purely logical studies of science by other kind of investigations that dealt with the non-logical aspects of the sciences, for instance, history, sociology and psychology of science. Whether this plan is convincing need not be discussed here. Rather, I'd like to point out that also in this more liberal conception of philosophy of science the monopoly of logic for the study of the formal structure of science remained intact. That is to say, the tool for dealing with the formal structure of scientific theories continued to be logic and logic alone.

This contention, however, gradually lost unanimous agreement. Van Fraassen hails Patrick Suppes to have been the first who envisaged another way of dealing with the formal structure of science by questioning the basic assumption of received philosophy of science. More precisely, he diagnoses the fundamental error as having put too much emphasis on matters linguistic:

> The mistake, I think, was to confuse a theory with the formulation of a theory in a particular language. The first to turn the tide was Patrick Suppes with his well-known slogan: the correct tool for philosophy of science is mathematics, *not* metamathematics. This happened in the 1950s – bewitched by the wonders of logic and the theory of meaning, few wanted to listen.[22]

In the decades after the 1950s the "semantic approach" in philosophy of science gained momentum. This is not to say that it could establish itself as a new unique orthodoxy. After all, Suppes's general recommendation to replace logic by mathematics as the basic tool of philosophy of science could be interpreted in many different ways – and indeed it was.

Suppes himself preferred a set-theoretical approach that conceived of scientific theories as set-theoretical structures. More precisely, he proposed to describe empirical theories such as mechanics or optics in terms of set-theoretical predicates in an analogous way as from a set-theoretical perspective a mathematical theory such as the theory of groups may be described in terms of the set-theoretical predicate "… is a group structure". Others, such as van Fraassen and Giere pre-

22 Van Fraassen, *Laws of Symmetry*, *op. cit*., pp. 221-222.

ferred a more geometrical account that reconstructed empirical theories essentially in terms of families of models or representations basically characterized by geometrical or topological structures, to wit, state spaces or phase spaces as the basic means for spatial representations in a generalized sense.[23]

Theories offer something like conceptual spaces, patterns for spatial activities, or maps. Having a theory is having a map that can be used to guide one's actions. Topology, as a general theory of space, investigates the structure of these generalized spaces. As long as one sticks to a narrow conception of (Euclidean) space this spatial characterization of a theory is doomed to remain vague and metaphorical, since clearly the "spaces" that are used in the various theoretical representations of science are not Euclidean ones. Here topology comes to the rescue, since it provides an ample spectrum of thoroughly analysized spatial concepts that can be used for this purpose.

5. Conclusion

Although in a short paper like this we had to leave out many details, it should have become clear that the philosophical vicissitudes of topology in 20th century philosophy of science offer a rich and multi-facetted agenda for history of philosophy of science that deserves further in-depth investigations. In particular the question why in the beginnings of the last century geometry lost its privileged status in philosophy and couldn't pass it on to topology requires further investigations. A too quick and simplistic answer would be that topology did not appear on the radar of philosophy because it was too technical and inaccessible a discipline for philosophers to squeeze some philosophical juice out of it. This answer is not convincing. Philosophy of science was often prepared to invest a lot of conceptual effort to come to terms with the intricacies of modern formal logic – and it is not always clear whether this was worth all the effort. Many examples show that philosophy of science did not shy away from considerable technical labor to come to terms with, say, quantum theory or relativity theory.

In contrast, philosophy of mathematics succumbed to the vice of elementarism or fundamentalism, as one may call it. Philosophers of mathematics, who subscribed to this doctrine located the philosophical relevance of mathematics entirely in its foundations, be they be claimed to be of a logical or set-theoretical nature or of any other kind. Consequently, topology and other advanced areas of mathematics disappeared from the agenda of philosophy of science. Such an attitude starkly contrasted with that of philosophy of the empirical sciences.

Recently the situation has changed again. After logic had lost its monopoly in philosophy of science, a new "mathematical philosophy of science" has begun to

23 Ronald Giere, *Explaining Science. A Cognitive Approach*. Chicago: The University of Chicago Press 1988, p. 20.

gain momentum. At least partially this new mathematical philosophy of science is informed by ideas that have a close affinity to geometry and topology. In other words, after having overcome the neglect from classical logic-centered philosophy of science the philosophical vicissitudes of topology and geometry continue to be an interesting topic on the agenda of history of philosophy of science.

Acknowledgment: Research for this this work is part of the research project FFI 2009-12882 funded by the Spanish Ministry of Science and Education.

Department of Logic and Philosophy of Science
University of the Basque Country (UPV/EHU)
P.O. Box 1249
20080, Donostia-San Sebastian
Spain
ylxmomot@ehu.es

Graham Stevens

Philosophy, Linguistics, and the Philosophy of Linguistics

Abstract

In this paper I suggest that, despite the overlap between philosophy of language and linguistics, philosophy of science has neglected linguistics. I argue that this has been to the detriment of philosophy of language. I examine the philosophical and linguistic treatments of definite descriptions as a case study to make this point.

1. Philosophy and linguistics

The claim that Linguistics is one of those sciences that received philosophy of science overlooked may strike some analytic philosophers as puzzling. After all, a large part of one of the core areas of contemporary analytic philosophy, namely philosophy of language, overlaps with theoretical linguistics to the extent that the two disciplines often appear indistinguishable. This is particularly true of work in the semantics and pragmatics of natural language. It is a somewhat arbitrary classification that decides whether David Kaplan's indexical semantics, or Paul Grice's systematisation of conversational implicature (to take just two examples among many) are contributions to philosophy or to linguistics. But this overlap, which I believe has been immensely fruitful, should not be mistaken for a philosophical interest in linguistics. It is evidence of a philosophical interest in linguistic phenomena, and one that has regularly led to a combining of forces among linguists and philosophers. But this interest in linguistic phenomena is quite distinct from an interest in the discipline of linguistics. Philosophy of linguistics is not philosophy of language (though, of course, there is common ground here too); philosophy of linguistics is perhaps best thought of as a branch of philosophy of science, akin to philosophy of biology, economics, or physics, etc. The philosophy of linguistics, in other words, is not so much interested in the subject matter of linguistics as it is in the status, nature and methodology of linguistics itself. Once the philosophy of linguistics is thought of in these terms, linguistics can quite understandably be labelled a science that received philosophy of science has overlooked.

Furthermore, it is very important to note that the overlap between philosophy of language and theoretical linguistics is an intersection of the subject matter of each that does not exhaust either. There are areas of the philosophy of language, such as the metaphysics of language (e.g. truthmaker theory, debates between

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_35,

Russellians and Fregeans over the metaphysics of propositions) that do not really touch on the concerns of linguistics, while only some of the concerns of linguistics have received the full attention of large numbers of philosophers. Philosophy of language has tended, even in those areas that do overlap with linguistics, to focus on issues concerning semantics and pragmatics. Syntax, on the other hand, has received far less attention from philosophers of language. This is decidedly different to linguistics where Chomsky's work has placed syntactic theory at the centre of linguistic theory. This difference has been exacerbated, as we will see shortly, by the inheritance of a notion of logical form in philosophy that is traditionally construed in radically different terms to the linguistic notion of logical form suggested by Chomsky that has become the orthodoxy in linguistics.

Syntax (at least the syntax of natural, as opposed to formal, languages) is a particularly striking example of an area of central interest to linguistics that has, until recently, been largely ignored by philosophy of language. There are many others, however: philosophy of language has, for example, paid no heed to central areas of linguistic enquiry like phonetics and morphology. Notwithstanding these important differences, however, there are general similarities between the two disciplines that should be mentioned. To some extent the general shape of ideological disputes that have dominated the two disciplines track one another: the ideological clash between the ordinary and ideal language philosophers (and more recently the contextualists versus the literalists) has its parallel in the clash between what are sometimes termed "essentialists" and "empiricists". Such general similarities, however, do not withstand closer scrutiny. For one thing, although terms like "empiricist" are often employed in linguistics, both as self-ascribed honorifics and as prejoratives for rivals according to the taste of the ascriber, it is doubtful that the term has the same force here as it does in philosophy. Whereas, in philosophy, empiricism is an ideology with a firm historical tradition founded on clear doctrines concerning what counts as providing adequate confirmation for beliefs and theories, and equally clear doctrines concerning the limitations of a priori enquiry, in linguistics the situation is different. In linguistics the most common ideological dispute in which the term gets employed is that concerning what counts as the genuine data of linguistic theory. Empiricists are those who reject linguistic intuitions as data, looking instead to actual uses of language for data. This opens up a fascinating methodological debate of great philosophical interest – particularly to the philosopher of linguistics – but it does not reliably track any disagreement between the empiricist and the rationalist in philosophy. Furthermore, actual linguistic research is not compelled to engage in such ideological dispute. Thus a given piece of research may well bridge this ideological divide in practice, with researchers employing mixed methodologies drawing on both approaches. This situation, for obvious reasons, is unlikely to be replicated in philosophy in the same way.

In the next section, I will look at a major point where philosophy of language and linguistics have, until very recently, been pursued largely independently of

one another, namely in their treatments of logical form. I will then focus on a particular example, that of the logical form of definite descriptions, to argue that this separation of the two disciplines has, in this particular case, been to the detriment of the philosophy of language. This will then support my overall conclusion that philosophy (of language in particular) has suffered as a consequence of philosophy of science's neglect of linguistics.

2. Logical form in philosophy and linguistics

Philosophers have inherited their conception of logical form from Russell. Though it was Frege who first recognized the importance of logical form, and of displaying it in a logical calculus distinct from ordinary language, it was Russell who embedded the concept of logical form in philosophy. Russell put a distinctly metaphysical spin on the notion of logical form. For example, in the following explanation of what he takes logical form to be, it is construed as the form of non-linguistic worldly items (in this case facts, though in earlier writings of Russell's they were construed as the forms of Russellian propositions[1]):

> The study of logic … is concerned with the analysis of logical *forms*, i.e. with the kinds of propositions that may occur, with the various types of facts, and with the classification of the constituents of facts.[2]

There are several reasons why Russell viewed logical form this way. Not least among them was his conviction that the neo-Hegelian philosophy that he was at pains to refute at the beginning of the twentieth-century was founded on an outdated logic that failed to recognize the importance of relations. This failure, Russell believed, was responsible for the idealism of the neo-Hegelians because it led them erroneously to the belief that it was contradictory to construe reality as something external to, and externally related to, the mind:

> Traditional logic, since it holds that all propositions have the subject-predicate form, is unable to admit the reality of relations: all relations, it maintains, must be reduced to properties of the apparently related terms … When once their reality is admitted, all logical grounds for supposing the world of sense to be illusory disappear.[3]

Despite the profound impact of this attack on neo-Hegelianism (it played a decisive role in establishing analytical philosophy as the dominant force in Western

1 For exegesis of Russell's metaphysics of propositions, see Graham Stevens, *The Russellian Origins of Analytical Philosophy*. London: Routledge 2005.

2 Bertrand Russell, "On Scientific Method in Philosophy" (1914), in: *Mysticism and Logic*. London: Routledge 1917.

3 Russell, *Our Knowledge of the External World*. London: Routledge 1914.

philosophy), it is not the doctrine of Russell's that has had the greatest impact on our understanding of logical form. The doctrine responsible for that was his much celebrated 1905 theory of descriptions, according to which denoting phrases are to be analysed as devices of quantification rather than reference, despite their shared grammatical role with devices of reference such as proper names.

Russell's denoting phrases are a subset of what linguists now call determiner phrases. These phrases are formed by attaching a determiner word such as "all", "some", "no", "every", "the", to a simple common noun like "book(s)", "table(s)", etc., or to complex nominal expression like "mountainous region(s)", "child(ren) of the Queen of England", "present King(s) of France", etc.

As just mentioned, Russell's list of denoting phrases is more restricted than the list of determiner phrases. All of those that Russell lists are definable in terms of the classical first order quantifiers "all" and "some" and the truth-functional connectives. He explicitly lists "all Fs", "every F", "some F", "a F", "any F", and "the F".[4] From what he says when explaining the merits of the theory, it is clear that he also includes possessive noun phrases like "France's present King", "my eldest son", and "Russell's theory of descriptions" among denoting phrases, on the grounds that they are semantically equivalent to definite descriptions, for example "France's present King" means the same thing as "the present King of France". Other determiner phrases were not considered by him, no doubt because he was unsure what to say about natural language quantifiers that resist expression in first-order (or even higher order) logic, like "most Fs", "few Fs" and "less Fs than the majority of Gs have heard of".

Other determiner phrases were excluded from the class of denoting phrases because Russell deemed them to be devices of reference, most notably complex demonstratives such as "that F". It is relevant here to note that this contradicts Russell's claim in "On Denoting" that denoting phrases are individuated purely in virtue of their form.[5] Complex demonstratives share the same form as Russell's denoting phrases but are treated differently because Russell takes demonstratives to be what he calls *logically proper names*, that is expressions whose sole semantic function is to refer. Without being diverted into Russell interpretation, we can make two observations about this. Firstly, it shows an ambivalent attitude towards language: Russell makes appeal to linguistic (grammatical) features of denoting phrases when it suits his case but sees no problem with flaunting these same appeals when it also suits his case; secondly, it shows that Russell's thinking about denoting phrases, despite his explicit claim that "a phrase is denoting solely in virtue of its form",[6] is really driven by semantic considerations. What places 'the present King of France' on the list of denoting phrases, and 'that person in the corner of the room' on the list of logically proper names is an intuition Russell

4 Russell, "On Denoting" (1905), reprinted in: Bertrand Russell, *Logic and Knowledge Essays 1901-1950*, ed. R.C. Marsh, London: Routledge 1956, p. 41.

5 *Ibid*.

6 *Ibid*.

has about the semantic contribution these expressions make, not reflection on their grammatical form. As we shall see in the following section, this is a clear example of where philosophy's lack of regard for linguistic data is to its detriment.[7]

In its clearest presentation, in Volume One of Whitehead and Russell's *Principia Mathematica*,[8] the theory of descriptions is presented by giving the logical forms of sentences containing denoting phrases in first-order logic. As Neale[9] has argued, it is not essential to the language that it be couched in this particular formal language.[10] The essence of the theory is that denoting phrases are devices of quantification, not reference. Though it is often thought to be primarily a thesis about definite descriptions, it is really a theory about all denoting phrases. Definite descriptions are simply the most controversial because the other denoting phrases are readily accepted by almost everyone as quantifier phrases. As a claim about definite descriptions, the central assertion of the theory boils down to the following claim:

> All phrases … containing the word *the* (in the singular) are incomplete symbols: they have a meaning in use but not in isolation. For "the author of Waverley" cannot mean the same as "Scott", or "Scott is the author of Waverley" would mean the same as "Scott is Scott", which it plainly does not; nor can "the author of Waverley" mean anything other than "Scott", or "Scott is the author of Waverley" would be false. Hence "the author of Waverley" means nothing.[11]

The theory that yields this result for Russell gives the following contextual definition for sentences of the form "the F is G":

(1) The F is G $=_{df} \exists x\,((\mathrm{F}x \wedge \forall y\,(\mathrm{F}y \supset x{=}y)) \wedge \mathrm{G}x)$

In other words, "The F is G", properly analysed, has the form "at least one thing *x* is F, everything is such that if it is F then it is identical with *x*, and *x* is G".

What exactly does this claim amount to? The *sentence* in question does not have the form given in (1). For example, the English sentence "The present King

---

7 For further discussion of the relation between complex demonstratives and Russell's denoting phrases, see Graham Stevens, *The Theory of Descriptions: Russell and the Philosophy of Language*. Basingstoke: Palgrave Macmillan 2011, Ch. 5.

8 Alfred North Whitehead and Bertrand Russell, *Principia Mathematica*. Cambridge: Cambridge University Press 1910–1913 in 3 vols.

9 Stephen Neale, *Descriptions*. Cambridge (Mass.): The MIT Press 1990.

10 Neale's claim has met with fierce opposition from some Russell scholars, most notably Bernard Linsky. See Bernard Linsky, "The Logical Form of Descriptions", in: *Dialogue*, XXXI, 1992, pp. 677-83 and "Russell's Logical Form, LF, and Truth-Conditions", in: G. Preyer and G. Peter, *Logical Form and Language*. Oxford: Clarendon Press 2002. See Stevens, *The Theory of Descriptions, op. cit*., Ch. 4 for a defence of Neale on this point.

11 Whitehead and Russell, *Principia Mathematica*, vol. 1, p. 67.

of France is bald", does not have the form of the (logico-) English sentence "at least one thing *x* is presently King of France, everything is such that if it is presently King of France then it is identical with *x*, and *x* is bald". It does not even have the same lexical components as this sentence, let along the same structure. The claim Russell is making is that the analysis reveals the form of the *meaning* of the sentence. That meaning, according to Russell is a non-linguistic abstract entity called a *proposition*. What Russell's theory is supposed to do therefore, is to reveal the proposition that is disguised in its linguistic representation. *Logical forms, in short, are the forms of propositions.*

3. The syntax of definite descriptions

The English sentence "The present King of France is bald" has the same syntax, taken at face value, as a sentence that contains a proper name in place of the definite description, such as "Russell is bald". Both combine a noun-phrase (hereafter NP) and a verb-phrase (hereafter VP) in the simple form [[NP][VP]] (we can ignore the internal structure of the VP):

(2) [[The present King of France$_{NP}$] [is bald$_{VP}$]]
(3) [[Russell$_{NP}$] [is bald$_{VP}$]]

We could note that the NP in (2) has a more complex internal structure than the NP in (3), for example by marking the determiner "the" or even by revealing the structure of the complex expression it attaches to. This will not be sufficient to reveal the structure that Russell attributes to the proposition expressed by (1), however. We can see this more clearly if we consider the negation of (1):

(4) The present King of France is not bald.

Anyone who has taught a course on the theory of descriptions to relatively theoretically unprejudiced undergraduates will know that intuitions clash over the truth-value of this sentence. Some see it as obviously false on the grounds that it affirms the existence of a present King of France, some as obviously true on the grounds that there is no present King of France who is bald. Russell's theory predicts and explains these competing intuitions, as it takes (4) to be ambiguous between the propositions whose forms are unambiguously displayed in the following two examples:

(5) $\exists x\,((\mathrm{F}x \wedge \forall y(\mathrm{F}y \supset x{=}y)) \wedge {\sim}\mathrm{G}x)$
(6) ${\sim}\,\exists x\,((\mathrm{F}x \wedge \forall y(\mathrm{F}y \supset x{=}y)) \wedge \mathrm{G}x)$

The difference between the two is that the description has wide scope in (5), narrow scope in (6), with regard to the negation operator. Yet this scope interaction is only possible if descriptions have the degree of logical complexity that Russell's account of them provides. No structure apparent in the ordinary sentence can support the distinction. Yet (5) and (6) reveal a genuine ambiguity in (4). The truth-conditions of (4) seemingly cannot be correctly identified if all we have to go on is its lexical components and grammatical structure.

If syntax is identified with the grammatical structures readily apparent in the sentence as ordinarily inscribed or uttered, there seems no choice but to follow Russell in looking for a non-linguistic source of this ambiguity. Chiefly due to the influence of Chomsky, however, linguistic theory does not standardly identify the syntax of a sentence with the grammatical form it displays in ordinary inscriptions or utterances. In Chomsky's generative project, a number of distinct levels of syntax are recognised. What we might pre-theoretically think syntax to be, is the surface structure (SS of the sentence). Information pertaining to SS is represented in (2) and (3). The sentence when uttered, has a phonetic form (PF) that may differ from its SS. Chomsky sees both levels of syntax as being derived systematically from a single underlying deep structure (DS). This structural level of syntax is taken to be uniform across languages and this plays a crucial role in Chomsky's postulation of an innate Universal Grammar. Discovering what the systematic processes are that act on DS representations to produce representations at other levels of syntactic representation is the primary task of linguistic theory. As we have already seen in considering Russell's theory of descriptions, however, we know that no theory of the structure of what we mean by our sentences will be adequate if it yields no greater structure than what is apparent in the sentence's SS or PF. More structure is needed if the grammar is to successfully map sentences to their truth-conditions. A further level of syntax is proposed for this role: the level of logical form (LF). LF structures contain all of the information relevant for semantic interpretation, such as quantifier scope, unarticulated pronouns, and so on. Unlike Russell, however, Chomsky maintains that LF is just another level of syntax – a genuine structural feature of the sentence. All these levels of syntax are taken to reside in a language faculty of the brain. LF structures are just representations of syntactic properties of the sentence that are not encoded in its SS or PF.

How do LF structures relate to Russellian logical forms? There has been some controversy surrounding this question. In his reconstruction of Russell's theory of descriptions, Stephen Neale moved noticeably towards a notion of logical form that was more in tune with Chomsky's LF. Although Neale did not fully commit his reconstruction of the theory of descriptions to a linguistic conception of logical form in *Descriptions*, he was close enough to it to draw criticism from those who wanted to defend a Russellian conception of logical form. Thus, Bernard Linsky objected to Neale's claim that he was defending a version of Russell's theory, on the following grounds:

> The lexical items to be inserted at the leaves of trees in logical form have significant ontological status for Russell. In contemporary LF they are simply the primitive words of the language. No assumption is made that each word corresponds to something in the world … I propose that it would be important to him to choose the right logical terms to be primitive, as they would have to reflect elements of logical form in the world: 'the' does not name any such constituent. Neale rejects these aspects of Russell's view, thinking that he is dropping something incidental to his purposes. Instead, the difference between a syntactically primitive lexical item and an item corresponding with a genuine constituent of the world is a large one, and not one to be glossed over.[12]

Whether or not Linsky was correct to accuse Neale of distorting Russell's theory by invoking a linguistic notion of logical form, he was certainly correct that Neale was invoking it. This is made more explicit in a subsequent paper of Neale's in which he defends the thesis that: "Arguably, we can make some serious progress by exploring the view that a fully worked out theory of LF will be a fully worked out theory of logical form".[13]

This is not the place to pursue the disagreement between Neale and Linsky over the exegesis of Russell. However, there is an important point to note about the transition to a linguistic notion of logical form of the sort that Neale recommends that is not discussed by either Neale or Linsky. If logical form is a syntactic structure, this will have significant consequences for Russell's arguments in support of the theory of descriptions. Consider, for example, the argument given above for the theory. Because (4) is ambiguous between the two readings whose truth-conditions are captured by (5) and (6), Russell argues that only if the propositions expressed by (4) have the structure described in (5) and (6) can they fit the predictions made by linguistics data. In other words, only a quantificational structure will account for the ambiguity because only it will allow for scope ambiguity. This argument, despite making appeal to syntactic/structural properties of the sentences/propositions, is a semantic argument. It is an argument that appeals directly to truth-conditional data in support of its account of logical form. Furthermore, logical forms just are determined by truth-conditions here. This is problematic, however, because the argument may be challenged on semantic grounds.

Gareth Evans noted in *The Varieties of Reference*[14] that if the semantic theory one adopts is that endorsed by the negative free logician, the claim that only quantifier phrases introduce scope phenomena into propositions when occurring in subject position because only they have the required syntactic complexity to do so is highly dubious. If a proper name *a* carries no existential import, then the sentence (7) will be ambiguous between the truth conditions given in (8) and (9):

---

12 Linsky, "Russell's Logical Form, LF, and Truth-Conditions", *op. cit*., p. 404.

13 Stephen Neale, "Grammatical Form, Logical Form, and Incomplete Symbols", in: A.D. Irvine and G.A. Wedeking (Eds.), *Russell and Analytic Philosophy*. Toronto: Toronto University Press 1993; reprinted in: G. Ostertag (Ed.), *Descriptions*: *A Reader*. Cambridge (Mass.): The MIT Press 1998, p. 95.

14 Gareth Evans, *The Varieties of Reference*. Oxford: Oxford University Press 1982.

(7) a is not F
(8) $\exists x(x = a \wedge \sim \mathrm{F}a)$
(9) $\sim\exists x(x = a \wedge \mathrm{F}a)$

Although (8) and (9) have a quantificational structure, this in no way means that (7) has a quantificational logical form. It is just that the semantic theory of free logic is one in which names lack existential import; only if an existentially quantified variable flanks an identity sign opposite to a proper name, do we have a formula carrying existential commitment. This gives a justification for construing names as having scope properties. We might represent these properties using a scope operator akin to that employed by Whitehead and Russell in *Principia Mathematica*:

(10) $[a] \sim \mathrm{F}a$
(11) $\sim\{[a]\ \mathrm{F}a\}$

In (10) a has wide scope, in (11) a has narrow scope. This possibility undermines what previously looked like a knock-down argument for the theory of descriptions. It shows that an alternative explanation for the competing readings of negated descriptive sentences is available that does not require attributing a quantificational logical form to descriptions. (10) and (11) show that NPs can have scope without being quantifiers. Therefore Russell's semantic argument is not sufficient to show that descriptions are quantificational.

A linguistic account of logical form along Chomskyian lines is not susceptible to this challenge. Descriptions are construed as quantificational on Chomsky's syntactic theory simply in the sense that, like other uncontroversially quantificational expressions, they are subject to syntactic operations such as movement, and binding. Purely syntactic data are appealed to in support of the claim that determiner phrases are moved in the transition from DS to SS and this movement process leaves a so-called *trace* which, while absent from the SS and PF is retained in the LF. This trace acts in just the same way that a bound variable does in the logical forms that Russell provided.

This is a striking case where philosophy of language has neglected linguistic theory to its own detriment. Philosophers of language have been engaged in a fierce debate over the semantics of definite descriptions for over half a century. Yet only recently have they looked to linguistics and discovered that it had an entirely new perspective on the debate that philosophers had neglected. Thanks to Neale and other philosopher of language who, like him, have raised awareness of the importance of linguistic theorizing to the philosophy of language, that situation has changed dramatically in recent years. It is still rare, however, to find philosophers seriously reassessing their conceptions of logical form in light of linguistic research once one looks beyond the philosophy of language.

4. Some conclusions

The neglect of linguistics is something that philosophers of language have recognized and begun to compensate for over the past decade or two. In relation to the particular example I have focused on in this paper, many philosophers of language now address this issue directly and an interest in the semantics/syntax interface is now commonly encountered, for example, in consideration of the hypothesis, endorsed by many philosophers of language, that LF is the interface between syntax and semantics – the level of syntax fit for semantic evaluation. This in turn raises rich and interesting questions about the relation between LFs and a hypothesised language of thought. Consideration of these sorts of issues puts linguistic theory at the heart of contemporary concerns in the philosophy of language. Similarly, the contextualist challenge to truth-conditional semantic theorising in natural language semantics has drawn philosophers of language and linguists into a shared debate about the semantics/pragmatics interface.

Despite these welcome points of overlap and interaction between the two disciplines, however, it is unclear that the philosophy *of* linguistics is significantly less neglected now than it has been over the past century. Very little has been written, for example, on what one might conceivably take to be the most fundamental question in the philosophy of linguistics: *is linguistics a science?* Yet the answer to this question may well shed important light on the questions that are occupying philosophers of language and linguists. The extent to which we should endorse a linguistic conception of logical form, for example, may well be judged in light of our answer to that question, and related questions about the proper methodology and data of linguistics. For the answer to those questions is likely to have significant implications for what we take syntactic theory, and linguistic theory in general, to reveal about the nature of language.

Philosophy, School of Social Sciences
University of Manchester
Oxford Road
Manchester
UK
graham.p.stevens@manchester.ac.uk

PSE Symposium at EPSA 2011: New Challenges to Philosophy of Science

Olav Gjelsvik

Philosophy as Interdisciplinary Research

Abstract

This paper raises issues about how philosophy ought to proceed. In the background are two competing approaches to the evidential grounding of philosophical insight. According to a widespread view, philosophical knowledge rests on a set of intuitions. According to another, philosophy has no special evidential grounding. This paper will resist the attractions of the first picture, and argue against the separateness of philosophy that it lends support. I shall try to make plausible that such a picture can be harmful both for philosophy and for empirical science. We should replace it with a mild form of unity of science.

1. Introduction

We have recently seen a development towards a philosophy of philosophy modelled on philosophy of biology, psychology and so forth. It is fair to say that the philosophy of science community has not contributed much to this development; it is rather a development by methodology-conscious philosophers with high general competence but no specific background in the philosophy of science. On the other hand there is, in this development, an acknowledgement of a desire to learn from the way philosophy of science has fruitfully approached various disciplines when turning to the discipline of philosophy.

The question of what philosophy is or ought to be is of course a hotly contested topic, more so than in almost any other discipline. One major aspect of that discussion concerns the evidential basis for philosophical knowledge (if there is such a thing). Disagreements on that point reflect substantive disagreement as to whether philosophy as a discipline is set apart from other disciplines, by for instance basing itself on *intuitions* and having conceptual insights as its aim, or whether there is no such separateness, and perhaps no such aim as (pure) conceptual insight. Alvin Goldman represents a mild form of the first line and writes, "One thing that distinguishes philosophical methodology from the methodology of the sciences is its extensive and avowed reliance on intuition."[1]

If we take the view expressed by Goldman literally, we see philosophy as very different from any empirical science, in that the evidential base is largely made up

1 Alvin Goldman, "Philosophical Intuitions: Their Target, their Source and their Epistemic Status", in: *Grazer Philosophische Studien* 4, 2007, pp. 1-26, p. 1.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_36, © Springer Science+Business Media Dordrecht 2013

of intuitions. Such a view raises a number of issues, not least issues about what intuitions are, how we know them, whether they carry their evidential status on the surface so to speak, in a manner not unlike Humean impressions or intuitions in ethics. A contrasting view would maintain that intuitions simply are theoretical judgements that can always be revised in the light of further theoretical development. In that case, they have no special evidential role. A third view could be quite agnostic about what intuitions are, and take no reductive stance towards intuitions, but argue case by case that intuitions play no interesting role in establishing philosophical knowledge.

This paper will resist the attractions of the first picture, and argue against the separateness of philosophy that it lends support. I shall try to make plausible that such a picture can be harmful both for philosophy and for empirical science. We should replace it with a mild form of unity of science.

2. Philosophical evidence and intuitions

One extremely general phenomenon that really lends some plausibility to the first picture that gives intuitions an important role, is our limited ability to provide fully informative accounts of general and philosophically central concepts, i.e. the concepts that philosophy is typically interested in, the concepts the exploration of which makes up the core of philosophy. Here we find concepts like justice, cause, knowledge, and explanation. One central theme in the history of philosophy of science has been the development of proper concepts/theories of explanation and also of cause. Take explanation as an example. Many theories/accounts of explanation have been refuted by counterexamples provided in the literature. We might conclude that the counterexamples show that a suggested account or theory of explanation lets in both too much and/or too little, and we hold that this fact refutes that theory/account in question. Thinking about how this insight is established, that the theory is refuted, the ground on which the insight rests, we might think that the counterexample elicits an intuition, and that this intuition is to be thought of as an insight into the concept of explanation that the suggested theory did not capture. When we thus intuit that relevant fact about the concept of explanation, we also realize by the same token that the suggested theory or account of explanation is refuted. Timothy Williamson has developed a careful alternative to this way of thinking in his discussion of the parallel issues in the case of knowledge. Still, many people here see a source for their way of thinking about intuition as the evidential base.

The questions Goldman's view gives rise to include questions about what intuitions actually are, and also whether there is, as a matter of fact, extensive reliance on them in actual philosophical practice. We can frame questions about the logic and semantics of the verb "to intuit" something in order to make progress

here, and do all this within the setting of seeing the core of philosophy as the exploration of the contents of a cluster of central concepts. Kirk Ludwig has contributed interestingly towards such an understanding of intuition.[2]

Intuitions, however, seem at least very close to judgements about conceptual content, i.e., in our case, the content of the concept of explanation. Whether they are judgements or not is of course the controversial point; if they are judgements they do not necessarily have the sort of special or exceptional epistemological role that they are given on Goldman's view.

There are various grand strategies of arguing against a view like Goldman's. One could argue like Quine against the very distinction between the analytic and the non-analytic. Timothy Williamson has argued for a view that is not different from Quine's regarding the relationship between the analytic and the non-analytic, and in doing so committed himself to far fewer controversial assumptions that Quine did.[3] Williamson argues both against the metaphysical version of view, i.e. the view that meanings or conceptual content make the analytical truths true, and also against the epistemological thesis that possessing or mastering a concept brings with it knowledge of analytical truths. If no interesting distinction between the analytic and the non-analytic can be upheld, then the picture of the core of philosophy as devoted to the exploration of conceptual truths as something very different from empirical truths cannot be upheld. If this is so, then there is no attraction in the view that the evidential ground on which philosophical knowledge rests is something special and very different from the ground of other kinds of knowledge.

Herman Cappelen's work on intuitions complements Williamson's work.[4] A main part of it consists in a careful and very valuable investigation into the question of whether philosophers do rely extensively on intuition even when they claim they do. His conclusion is that they do not; they do not rely on intuition in their work in any other sense than that of appealing to theoretical judgements. There is thus a mismatch between what many philosophers do, and their own understanding of how they go about it.

3. How ought philosophy to proceed?

Both Williamson and Cappelen seem to me to be basically right in their claims. Still there is an issue about how far they take us in the direction of any normative issue about how philosophy ought to be conducted. Williamson's basic message is that many philosophers' basic conception of their own activity is faulty: they are

2 Kirk Ludwig, "Intuitions and Relativity", in: *Philosophical Psychology* 23, 4, 2010, pp. 427-445.

3 Timothy Williamson, *The Philosophy of Philosophy*. Oxford: Blackwell 2007.

4 Herman Cappelen, *Philosophy without Intuitions*. Oxford: Oxford University Press 2012.

trying to do something which is in fact impossible when they attempt to achieve insight into the conceptual side of a terrain marked by a sharp distinction between conceptual truths and empirical truths. That sharp distinction has been with us from Hume and Kant, and is a present and powerful force. Cappelen's basic message is that philosophers don't do what Goldman claims they do or what they often claim themselves to be doing; they do not in fact rely on intuitions in their arguments or in the way they try to establish their philosophical theses. This is good news if Williamson is right; they are not trying to do the impossible after all even if they think they are. Still little follows directly from Williamson's and Cappelen's writings about 'ought' issues: whether philosophy is in a happy state and ought to follow its present course, or whether it would benefit from clear changes in the way it is conducted. They have, however, prepared the ground for possible changes. As long as there is no special evidential ground that philosophical knowledge rests upon, then what is good for philosophy is simply to answer its questions and do that in the best possible way.

It is within this general perspective that I want to present some specific aspects of the philosophical work on addiction. This is much less of a jump than it might seem at first. The basic lesson I want to draw is that taking in theoretical/empirical results in this area can be surprisingly fruitful for a range of deep philosophical questions that have been with us for centuries, questions that look like conceptual questions. I also want to claim that philosophical engagement with the empirical sciences on an issue like this can be fruitful for how the empirical sciences are conducted, what questions they ask etc. If this is right, it provides some ground for an ought: philosophy ought to concern itself more directly with at least some specific issues in the sciences; that it is good both for philosophy and for the sciences. Progress in philosophy as in many other fields comes in surprising ways. The philosophy of science has examples of a similar sort, but let us focus on this one case.

4. The significance of addiction

Addicts pop up in philosophical discussions, and the concept of addiction has now received quite a bit of philosophical interest. The first example I will point to is Harry Frankfurt's 1971 discussion of freedom of the will.[5] The addict is there portrayed as someone compelled to consume the addictive substance no matter what. There is 'total' absence of freedom of the will, and the compulsion is so conceived that compelled consumption cannot really be intentional action. Further research has shown that Frankfurt's conception of an addict is much too simplistic, but for the moment we can ignore that point. We have two dimensions in play by

5 Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person", in: *The Journal of Philosophy* 68, 1971, pp. 5-20. Reprinted many places.

bringing addiction into the issues that Frankfurt was working on, or two clusters of issues, one around the concept of freedom of the will or autonomy, and the related concept of responsibility, and also one around the concept of intentional action. The issues around intentional action hook up with conceptions of weakness of the will (or acrasia). Let me also add that it might of course have been a mistake of Frankfurt to bring in the case of addiction, given his purposes. I shall maintain that it was not. If there was a mistake it was rather that he did not take such phenomena as addiction sufficiently seriously when trying to understand freedom. There is undoubtedly some impairment in freedom of the will when you are addicted to heroin or alcohol, even if the impairment is very different from the way Frankfurt conceived of it. The type of impairment throws important and unexpected light on what freedom of the will is.

There is more. The case of addiction is clearly an area where different scientific disciplines approach the same phenomenon. This is of great interest for many reasons, and it raises, in complex ways, questions about the relationships between the questions these disciplines try to answer. On the one hand we have the decision sciences, and economics, and their ways of providing accounts of addiction, as rational choice in some cases, or as the result of the operation of mechanisms of irrationality understandable only within a choice framework, like hyperbolic discounting, or similar such things. People within these sciences mainly interact with other members of their own discipline. In the biological models of addiction there is a tendency to say this is addiction, and then identify some more or less permanent change in the mesolimbic system or in nucleus accumbens. As it turns out, such permanent neurophysiological changes are far from perfectly correlated with the behavioural patterns that the decision sciences attempt to explain with their approaches; some things thus seem to be better explained by the latter, and knowledge is lost if the whole focus is the former. The relationship between the neurological modelling and the decision-theoretic modelling is, therefore, delicate and challenging, and also marked by limited interaction between the disciplines.

One of the most important results from the addiction research is the basic finding of Berridge and Robinson about the structure of the motivational system.[6] They claim to have found that motivation in mammals can be seen as two cooperating but dual systems, one system to be thought of as expressed by "wantings", the other by "likings". Likings seem to be cognitively informed, and to be stable the way being cognitively informed is stable. It therefore has many properties in common with judging something to be best, while at the same time it seems to have some

6 Kent C. Berridge, and Terry E. Robinson, "The Mind of an Addicted Brain: Neural Sensitization of Wanting versus Liking", in: *Current Directions in Psychological Science* 4, 1995, pp. 71-76; Kent C. Berridge, and Terry E. Robinson, "The Psychology and Neurobiology of Addiction: An Incentive–Sensitization View", in: *Addiction* 95, Supplement 2, 2000, pp. S91-S117; Terry E. Robinson, and Kent C. Berridge, "The Neural Basis of Drug Craving: An Incentive-Sensitization Theory of Addiction", in: *Brain Research Reviews* 18, 1993, pp. 247-291.

motivational elements (this is, however in need of further investigation, and seems somewhat undetermined by data). "Wantings" behave differently, and are directly motivational. The picture is, furthermore, that these two systems normally march in step. The important bit is that they can come apart in special cases where this normal working of the motivational system of the organism is suspended. In those more exceptional circumstances we see cases of liking something and not wanting it, and wanting something and not liking it, to the point of disliking it. Wantings can fluctuate in dramatic ways when circumstances are special. More specifically, being addicted to addictive substances can influence mammals to desperately want things, i.e. wanting to consume a substance, even when they do not like doing it. It is still fair to say that a motivational monism can get pretty far in explaining many of the same things as the dual approach of Berridge and Robinson, and that has been done by introducing mechanisms like hyperbolic discounting in the fashion of George Ainslie.[7] The fact that there seems to be neurophysiological evidence in favour of the dual approach is extremely interesting and challenging, and judging between these approaches requires a very balanced theoretical judgement that takes in very many concerns, maybe also philosophical concerns.

Let me just stop and take stock at this point and say something about how the findings in this area can be of great help when resolving philosophical issues, and also something about how philosophy can contribute to the ongoing research that should result. I will take the last issue first, even if this order might seem to be getting things the wrong way round.

5. Benefits from philosophy to ongoing empirical research

a) The basic issue of whether the general approach to motivation should be monist or not is an old one in philosophy, and in the philosophical tradition there has been a number of ways the oppositions between such views have been played out. This background seems fruitful and important for the full assessment of the theoretical conflict between dual and monist approaches to motivation. Philosophy can therefore, in this area, be an important contributor to generate hypotheses that can be empirically tested in ways not conceived of in earlier days.

b) Philosophy and philosophy of science can contribute in a number of ways in clarifying and resolving some of the issues around very different disciplines trying to address and explain what looks like the same phenomenon. This is indeed one of the specialities of philosophy: to analyse and settle whether explananda are the same or not, and whether explanations are competing explanations or not. Against this background, philosophy can contribute to the theoretical judgements required for answering whether dual and monist approaches to motivation are indeed competing. There is here a three-way meeting point between neuroscience,

7 George Ainslie, *Picoeconomics*. Cambridge: CUP 1992, and many later writings.

behavioural science and philosophy. The best way forward seems to be enlightenment about what the various disciplines and approaches try to do and how they do it.

c) Philosophy, and the philosophy of science as well, can contribute towards seeing and working out the more general significance of some of the findings in the more empirical sciences. The particular benefits might be to work out a new and scientifically informed conception of agency with an informed focus towards how to think of impaired agency, and look at how such a conception matters for the required theoretical judgements about what is impaired in what way, what is not impaired etc. That is now in the process of being done by Richard Holton and Kent Berridge, and of course by many others.[8] Much of this work will have major repercussions for how to do philosophy of action.

Against the background of such work one might also hope for an account of addiction that clearly exhibits how freedom gets impaired when the addiction develops, and a full theory of addiction might be getting closer. That will bring with it possibilities for major progress for how society deals with addictions, how to think of addicts' responsibilities for their acts, whether they are able to give informed consent for various types of treatments and so on.

6. Benefits to philosophy from empirical research

The possibility of a dual approach to motivation grounded in neurophysiology can be of great help when trying to make progress on some very deep issues in philosophy. Here are some examples:

a) The debate about whether normative or moral judgements are intrinsically motivating (internalism/externalism issues) ought to be informed by such new findings. Much of the discussion in philosophy is simply blind to the possibilities which parts of empirical science now seem to entertain, and very many of the arguments put forward in philosophy will have to be reconsidered in the light of the possibility that we can fail to want things we like (or judge best), and we can like (or judge best) things we do not really want. Traditional justifications of internalism and externalism in philosophy are potentially all in trouble. Let me add that there are further delicate issues here about how close "likings" are to judgements about what is best or right to do in an agent capable of such judgements. The interactions with internalism and externalism issues might actually be a way towards much greater clarity in how to think about likings.

b) How to conceive of acrasia or weakness of the will have to be discussed all over again. This cluster of problems makes up a very old and deep set of issues in philosophy. Many of the arguments against the possibility of acratic action, or weak willed action, from ancient time will have to be reconsidered, partly because

8 The work Holton and Berridge are doing together is not published yet.

they rely on specific internalist conceptions of normative judgement we perhaps cannot uphold in the light of the present findings and the support for the distinction between wanting and liking. There is some vague parallel between the present findings and old theses about partitioning of the soul, and these issues have to be further explored as well. They might throw much light on how we should relate to such thoughts.

c) We also have to reconsider philosophical accounts of compulsion that seem to block conceiving of compulsion as intentional action with some, albeit limited, sort of freedom. This brings in a full range of issues about the extension of the concept of doing something intentionally, what that is, and what the things are that we do intentionally, and the relationships between what we do intentionally and what we are free to do.

d) Another whole area in philosophy that can benefit much from this new knowledge is all the work around the mind-body problem. The various types of dynamic interaction between the neurophysiological level and the intentional level in addictive behaviours is a rich source for new and more concrete ways of thinking about this age-old problem.

e) A further area has to do with developing informed views on addiction which can ground both ethical and legal decisions involving addiction, responsibility of addicts etc. The question I raised above, about whether heroin addicts can give informed consent to be given free heroin in experimental treatment, is a burning question in today's society, and we need, as a society, to be better informed than we are to be able to answer that question.

Final thought: the background outlined seems to me to make possible new and constructive interactions between work in more general philosophy and more specialized work in the philosophy of science, to the mutual benefit of both. The focus on phenomena that are being studied from different disciplines and where the findings can have major implications for philosophy itself seems to be the exactly right focus to bring about new and interesting interactions between philosophy of science and philosophy. To my mind that interaction has suffered lately. Here is a way forward with benefits to all.

7. Concluding remarks and a normative perspective

I set out to provide reasons for a change in the way philosophy is done towards a certain type of interdisciplinarity; towards learning from the empirical sciences in ways relevant both for making philosophical progress and for laying the ground for very constructive interaction between philosophy and the these other disciplines, for the benefit both for all disciplines involved and for society at large.

There is a large number of reasons in support of a quite specific ought. This ought directs towards learning from science, and integrating that learning into the

way we do philosophy, instead of trying to elicit the most sophisticated intuitions from armchair philosophy. In many of the areas described above, philosophy proceeds in a traditional way without paying almost any attention to work in various sciences that can actually benefit philosophy, and attempts at actual interactions are at best very scarce. The philosophical work would still mainly be conducted in the same armchair even after a change of ways, but not by calling up intuitions that are to be had in splendid isolation from knowledge of empirical science. It would have to be pursued by two-way interaction with a large number of other sciences and disciplines, with philosophy situated at the theoretical end of the mutually enlightening work. That provides the ground for a good way of doing both many parts of philosophy and some parts of science, and thus provides grounds for an ought-statement to the same effect. If this ought I am defending is a fact, that fact seems fully compatible with seeing intuitions as theoretical judgements, and, on the other hand, it is in some clear tension with seeing intuitions as some sort of (independent) evidence.

It is probably fair to say that I have picked an area which is ideal for support for the conclusion I am drawing about how philosophy ought to proceed. Even if this is so, it does not detract essentially from the point that in many areas of philosophy progress is least likely to come from continued traditional efforts where philosophical issues are discussed in complete isolation from relevant neighbouring sciences. These things will actually be somewhat different in different parts of philosophy, but the point here is that concepts/discussions about things like blameworthiness, responsibility, doing something intentionally, being free to do something, internalism versus externalism in approaches to motivation, evaluative concepts etc, are central philosophical issues and have been so for centuries. There can be other parts of philosophy where gains from interdisciplinarity are much smaller than the case of addiction and the related cases. That is, so far, just an open issue. I believe there are many cases that support the same general type of conclusion as I have drawn here, but I limit myself to this one.[9]

CSMN
Department of Philosophy, Classics,
History of Art and Ideas
Faculty of Humanities
University of Oslo
Box 1020 Blindern
0317, Oslo
Norway
olav.gjelsvik@csmn.uio.no

9 I am very grateful to Nick Allott for comments.

Theo A. F. Kuipers

Philosophy of Design Research

Abstract

The paper investigates the conceptual possibility of a threefold distinction in design research that parallels the threefold distinction of laws, theories and research programs in nomic research, viz. design laws, design theories, and design research programs. In view of a rather different picture of design or, more broadly, applied theories of Ilkka Niiniluoto, the paper leaves the challenge to start comparative case studies to evaluate the two conceptions.

1. Introduction

In his editorial introduction to *Philosophy of Technology and Engineering Sciences*, Volume 9 of the *Handbook of the Philosophy of Science*, Anthonie Meijers wrote:

> Not so very long ago most philosophers of science maintained that the subject-matter of this volume was uninteresting […] because technology was taken to be an applied science in which the application itself presented no new philosophical challenges.[1]

The message of Meijers is very true, even in the much wider sense of disinterest in applied and design sciences in general. Sure, there have been exceptions in the past, e.g. von Wright[2], Bunge[3], and Simon[4]. However, witness the volume referred to, containing 41 contributions, there is a growing interest in the last two decades, and rightly so.

1 Anthonie Meijers, "General Introduction", in: Anthonie Meijers (Ed.), *Philosophy of Technology and Engineering Sciences*, *Handbook of the Philosophy of Science*, vol. 9. Series editors Dov Gabbay, Paul Thagard and John Woods. Amsterdam: Elsevier 2009, pp. 1-19, p. 1.

2 Georg Henrik von Wright, *Norm and Action*. London: Routledge and Kegan Paul 1963.

3 Mario Bunge, "Technology as Applied Science", in: *Technology and Culture* 7, 1966, pp. 329-349. Reprinted in Friederich Rapp (Ed.), *Contributions to a Philosophy of Technology*. Dordrecht: Reidel 1974, pp. 19-36.

4 Herbert Simon, *The Sciences of the Artificial*. Cambridge (Mass.): The MIT Press 1969 (2nd ed. 1982).

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_37,

On the question "What is technological science?" Sven Ove Hansson[5] discusses six defining characteristics that distinguish technological from the other (natural) sciences:

(1) they have *human-made* rather than natural *objects* as their (ultimate) study objects,
(2) they include the practice of engineering *design*,
(3) they define their study objects in *functional terms*,
(4) they evaluate these study objects with category-specific *value statements* (safety, health),
(5) they employ less *far-reaching idealizations* than the natural sciences,
(6) they have no need of an exact mathematical solution when a sufficiently close approximation is available.

In combination, the six characteristics are sufficient to show that the technological sciences are neither branches nor applications of the natural sciences, but form a different group of sciences with specific characteristics of their own.

The focus of this paper is the broad notion of science-based design research or, simply, design research. Note that speaking of design *sciences* would be misleading because in modern science there is design research in almost all academic disciplines. Talking about descriptive and explanatory sciences is for the same reason misleading.

I will discuss design research on three levels, suggested by the level distinction of (nomic) laws, theories, and research programs that turns out to provide an illuminating viewpoint in the area of theory oriented descriptive and explanatory research, here briefly called nomic research.[6] By analogy, three levels of design research can be distinguished:

1) technical norms[7, 8], or, as I call my version, design laws,
2) design theories, to be identified by the analogy,
3) design research programs[9].

5 Sven Ove Hansson, "What is Technological Science?", in: *Studies in History and Philosophy of Science* 38, 2007, pp. 523–527.

6 Theo Kuipers, "Laws, Theories, and Research Programmes", in: Theo Kuipers (Ed.), *Philosophy of Science: Focal Issues, Handbook of the Philosophy of Science,* vol 1. Series editors Dov Gabbay, Paul Thagard and John Woods. Amsterdam: Elsevier 2007, pp. 1-95.

7 Von Wright, *Norm and Action*, *Ibid*.

8 Ilkka Niiniluoto, "The Aim and Structure of Applied Research", in: *Erkenntnis* 38, 1, 1993, pp 1-21. Niiniluoto, "Approximation in Applied Science", in: Martti Kuokkanen (Ed.), *Idealization VII: Structuralism, Idealization and Approximation*, *Poznan Studies in the Philosophy of the Sciences and the Humanities*, vol. 42. Amsterdam: Rodopi 1994, pp. 127-139.

9 Theo Kuipers, Rein Vos and Hauke Sie, "Design Research Programs and the Logic of

In Section 2 I will briefly discuss design laws by starting from Ilkka Niiniluoto's view on technical norms, followed by an attempt to improve his characterization by introducing the distinction between structural and functional properties. In the longer Section 3 design theories will be defined as a kind of generalization of design laws. As indicated above, design theories will be identified on the basis of the analogy with nomic research. Hence, design theories look a bit like descriptive and explanatory theories. To be sure, the idea of design theories is not new. However, as we will see, its most clear explication in the literature I know of is that of Niiniluoto.[10] He claims that theories in what he calls applied sciences, including sciences in which design research is dominant, have a structure that is more or less the opposite of theories in basic sciences. In Section 4 I will, based upon earlier work, briefly discuss the nature of design research programs and indicate similarities and differences with nomic research programs. Some conclusions and suggestions follow in Section 5.

2. Design laws

Following von Wright[11], Ilkka Niiniluoto[12] views design research as a kind of applied research that serves epistemic and practical utilities and focuses on the establishment of technical (or practical) norms. The strongest version of such norms reads in the 1994-version:

"If you want *A*, and (you believe) you are in a situation *B*, then you ought to do *X*."

He argues that practical norms have a truth value, which enables support by basic research. The norm is true if and only if the corresponding causal statement "X causes A in situation B" is true.

Although all this makes much sense, it is in my opinion illuminating to introduce a distinction that is also frequently used in biology, notably when functional claims or laws are at stake. For example: the function of heartbeats is the circulation of blood in the body. I mean the distinction between structural and functional properties (dispositions, events, processes, etc.). In general, the distinction is a relative one, that is, a functional property on one level can be a structural property on a higher level of complexity. Moreover, both types of properties are observable, at least as soon as the properties are explicitly named. In the context of design research the distinction deals more specifically with properties that can be directly

Their Development", in: *Erkenntnis* 37, 1, 1992, pp. 37-63. Theo Kuipers, *Structures in Science*. Dordrecht: Kluwer 2001. Chapter 10.

10 Niiniluoto, "Approximation in Applied Science", *loc. cit*., Section 1.

11 Von Wright, *Norm and Action*, *Ibid*.

12 Niiniluoto, "The Aim and Structure of Applied Research", *loc. cit*. Niiniluoto, "Approximation in Applied Science", *loc.cit*.

imposed by humans on the one hand and properties that can only indirectly be aimed at on the other. Hence, it is an artificial version of the distinction in biology, where in the course of evolution functional properties have indirectly been realized by the arising and changing of structural properties.

From this perspective, it is as plausible to talk about design laws as about functional laws in biology and hence to define a design law as a true causal (observational) law or regularity of the form:

> Functional property A in situation B can be achieved by imposing structural property X.

Or, in the symbols and phrasing that I prefer for the general version:

> Imposing structural property S in context C causes functional property F.

A typical example is:

> Imposing bumps (S) at a certain place (C) causes a reduction of the car traffic speed (F).

Hence, a design law is a special kind of (observational) causal law.

Comparing Niiniluoto's technical norms with our design laws in terms of structural and functional properties, they seem two compatible ways of expressing means – ends relations. Both explications reduce these statements to causal regularities. Whereas functions can already be discussed in biological systems even when there is no agent making design, technical norms explicitly involve actions (do X) and agent causality, and design laws do so implicitly by the phrase 'imposing structural properties'. However, as Niiniluoto continued his comparative remark in correspondence, design laws do not include the idea of oughts, which is important to von Wright's technical norms and to Simon's 'sciences of the artificial'. Be this as it may, I have always been reserved about ought, or necessary condition, terminology, also in the context of intentional and functional explanation,[13] because it tends to hide the possibility of functional equivalents.

3. Design theories

To represent, usually very complex, cases of design research, e.g. drug design, speech technology, food technology, traffic technology, one needs compound representations. The aim of (complex) design research can be adequately represented within a space of (possibly) relevant properties by property profiles of (prototypes of) the products to be made and by two, mutually orthogonal, distinctions:[14]

13 Kuipers, *Structures in Science*, *loc. cit*. Chapter 4.

14 Kuipers et al., "Design Research Programs and the Logic of Their Development", *Ibid*. Kuipers, *Structures in Science, loc. cit*., Chapter 10.

- structural and functional profiles,
- desired and operational profiles.

For the first distinction it is presupposed that the space of relevant properties can be subdivided into a functional and a structural subspace. A functional profile collects the total of functional properties of a (potential) product, thereby implying that it lacks the other properties in the subspace of functional properties. Similarly, a structural profile collects the total of structural properties of a (potential) product, thereby implying that it lacks the other properties in the subspace of structural properties.

A desired profile collects the total of wished for properties of an intended product, thereby implying that the other properties in the space of properties are undesired or just optional. An operational profile collects the total of actual properties of (a prototype of) a product, thereby implying that the other properties in the space of properties are absent.

In combination, and assuming that there is already (a prototype of) a product, say x, this leads to four kinds of twofold profiles:

- a desired functional profile (DF) and a desired structural profile (DS),
- an operational functional profile (OF(x)) and an operational structural profile (OS(x)).

On this basis, design theories can be formulated, tested, and improved. A design theory presupposes the above distinctions and deals with claims of the following type, leaving a contextual reference implicit:

> Functional profile F can be achieved by structural profile S, or as I prefer,
> Imposing structural profile S causes functional profile F.

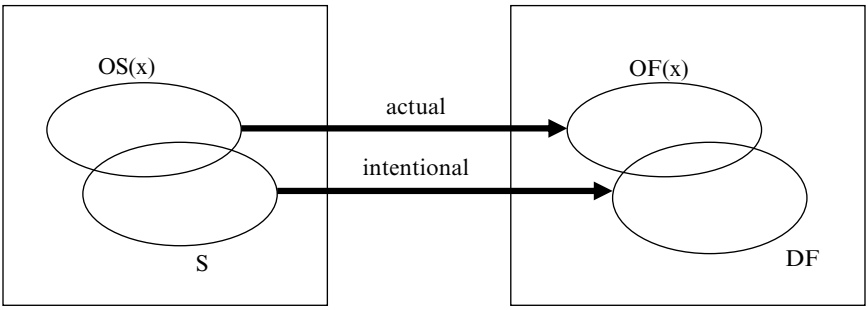
This type of claim has two versions, depending on whether it refers to an existing or an intended artefact:

Actual: Operational structural profile OS(x) causes (operational) functional profile OF(x).

Intentional: Imposing structural profile S will cause desired functional profile DF; formally: for all x, if S(x) then OF(x)=DF.

I define a design theory more specifically as a tuple of the form <S, DF> with the corresponding intentional claim, which is further on simply called the claim of the theory. A design theory is typically science-based when its claim is based on theoretical considerations, here called the underlying nomic theories, which make testing of the design theory, see below, also testing of these underlying theories.

The following picture characterizes the problem state of design research at a certain moment, assuming that there is already a prototype (x).



A problem state of design research

Legend:
OS(x)/OF(x) operational structural/functional profile of prototype x
S/DF structural profile/desired functional profile
→ causal relation

Testing the claim of a design theory <S, DF> is of course done by making a prototype, x, having the structural profile S and checking whether OF(x) = DF holds. Two possibilities should be distinguished. As the picture already suggests, the prototype may or may not be such that OS(x) = S. Let us start with the second case, OS(x) ≠ S in which case the theory does not claim that OF(x) = DF. However, assuming the theory is true, it is plausible to try to make progress by making a new prototype having a structural profile that is more similar to S and, due to the theory's truth, having a functional profile that is more similar to DF. The option hinges upon the heuristic default principle that more similar structural profiles cause more similar functional profiles. Another option appears when, for some reason or other, OF(x) turns out to be at least as attractive as DF. This opens the way to adapting DF in the direction of the OF(x), which occurs for example quite often in pharmaceutical research, as pointed out by Rein Vos in his aptly called book *Drugs Looking for Diseases* [15].

In the first case, OS(x) = S, the theory claims OF(x) = DF. If the latter claim is true, the theory is confirmed. Of course, this does not yet prove that the general claim of the theory is true. Hence, the prototype will have to be reproduced and tested again. However, if the claim OF(x) = DF turns out to be false, the theory is falsified. Assuming DF fixed, the task is of course to revise the theory by revising S. Another option we have already met: also in this case, OF(x) may be at least as attractive as DF.

In the next section we will return to the suggested moves in the context of design research *programs*, but here we will focus on similarities and differences between design and nomic theories.

15 Rein Vos, *Drugs Looking for Diseases*. Dordrecht: Kluwer 1991.

The (causal) claim of a design theory, recall: imposing structural profile S causes functional profile F, resembles the (realist) claim of a nomic *proper* theory, viz. that the theoretical entities and properties cause and explain such and such observational facts. Typical examples of explanatory causal proper theories are, e.g. the theories of Newton and Darwin, and utility theory. Another important similarity[16] is the following:

> Design research aims at (theories about) products with certain desired functional features.
> Nomic research aims at theories about phenomena with certain (desired) observational features.

However, there are also important differences:

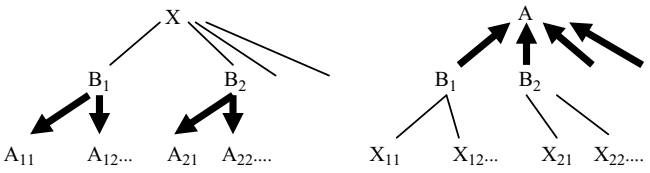
1) theoretical entities and properties are not observable,
   structural properties are, like functional properties, observable,
2) observational features are derivable from the proper theory,
   functional properties are (claimed to be) caused by structural properties.

Hence, whereas nomic proper theories explain observational features by non-observational ones, design theories are a kind of descriptive causal theories, similar to e.g. complex chemical reaction equations. Of course, if the causal relations are not transparent or not based on well-established theories, they ask for explanation of the working of the products. A typical example is the research around the working of aspirin.

The suggested concept of design theories is rather different from Niiniluoto's proposal.[17] He contrasts theories in applied sciences, covering predictive and design sciences, with theories in the basic sciences. Let me call them applied versus basic theories. In a basic theory there is one causal factor X, for example gravity, that is applied in a variety of situations B_1, B_2, … with the resulting effects A_{11}, A_{12}, …, A_{21}, … Hence, a basic theory is a collection of specific causal statements united by the same causal factor. In contrast, an applied theory specifies various means X_{11}, X_{12}, …, X_{21}, … for obtaining in different situations B_1, B_2, … a single goal A. Hence, it is a collection of technical norms (design laws) united by the same intended effect. The goal may be more or less specific. Very general goals are, for example, health and peace. In a picture, adapted from Niiniluoto, we get:

16 Kuipers, *Structures in Science*, *loc. cit*., Chapter 9.

17 Niiniluoto, "Approximation in Applied Science", *loc.cit*.



A basic theory An applied theory

In my view, nomic and design theories are similar: complex entities ideally generating desired observational features and desired functional features. Of course, there is no reason to choose between the two explications of applied/design theories. Both may be useful. It is just important to distinguish them.

The picture suggests perhaps a somewhat different interpretation. A basic theory as depicted may be conceived of as a theory-net in the structuralist sense, such that each triple $<X, B_i, A_{ij}>$ represents a theory-element. Of course, it is then plausible to construe an applied theory also as a theory-net, with theory-elements. What remains, however, is that they have an opposite structure.

Another possibility, suggested by Niiniluoto in correspondence, might be to see an applied theory as a special case of a design theory where the desired functional profile can be expressed by just one property A, e.g. health, welfare, etc. This would require e.g. illness dependent structural profiles and lead to another kind of theory-net. Indeed, it seems that the two explications would become compatible in this way.

4. Design research programs

Finally, the development of a design research program, assumed to be guided by some core idea, the so-called 'lead', can be described as a sequence of (or even a network of) changing profiles.[18] This enables to easily indicate similarities and differences with nomic research, notably their respective kinds of progress, to which we will return soon. Moreover, the cooperation between a design program and one or more other research programs can be described, in the apt terms of Henk Zandvoort, as that between a guide program and one or more supply programs.[19] On its turn, this can be seen as an explication of the main ideas of the Starnberger Gruppe in the 70s, e.g. about aerodynamics.

I noted already two similarities between design and nomic research and theories, viz. their aims and their claims, respectively. And I alluded already to fur-

18 Kuipers, *Structures in Science*, *loc. cit*., Chapters 1 and 10.

19 Henk Zandvoort, *Models of Scientific Development and the Case of NMR*, *Synthese Library* 184. Dordrecht: Reidel 1986.

ther similarities in the context of research programs. The main similarity is that between their kinds of progress. Progress in a design research program amounts to: the new prototype possesses more desired functional properties and fewer undesired ones than the old prototype. Progress in a nomic research program can analogously be characterized in terms of an increase of desired features and a decrease of undesired features.[20] Moreover, in both cases, a particular way of making progress is by the method of idealization and concretization. One starts with a prototype or theory that implements the core idea of the solution of the problem, but that neglects a number of relevant features. In successive stages it is attempted to build such desired features in the prototype or theory, and similarly for the elimination of undesired features.

Another similarity is that between the existence of functionally equivalent prototypes (and products) in design research and of observationally equivalent nomic theories. Such functional equivalents enable the arising of competing research programs in both types of research.

But there are also important differences. I noted already some of them in terms of design versus nomic proper theories: observable versus non-observable properties and descriptive versus explanatory claims. Other differences that come best in view in terms of research programs are:

1) a desired profile, or design target set, is a free and changeable choice, whereas given a domain and a vocabulary, the nomic target set is fixed;
2) a design target set is known in principle, whereas the nomic target set typically is not;
3) prototypes can straightforwardly be compared (comparative evaluation) both with regard to structural and functional properties, in contrast to the necessarily indirect evaluation of the theoretical part of theories, viz. by comparing their observational consequences;
4) to revise a prototype may be expensive for material reasons, to revise a theory is relatively inexpensive;
5) for a fixed domain and vocabulary, nomic truth approximation is, ideally speaking, free from external influences, whereas design research is basically open to such influences.

Despite these differences, there is in general the feeling that design and nomic research have much in common. This is strengthened in case the actual research is in fact a cooperation between a guiding design research program and one or more nomic programs supplying problem solutions, the paradigm case of science-based design research.

20 Kuipers, *Structures in Science*, *loc. cit*., Chapter 9.

5. Conclusion

In this paper I have shown that it is plausible to introduce a threefold distinction in design research that parallels the threefold distinction of laws, theories and research programs in nomic research, viz. design laws, design theories, and design research programs. The resulting picture of, in particular, design theories is rather different from that of Niiniluoto, who construes applied theories, including design theories, as a kind of opposite of basic theories. At face value, both constructs may have their value. Case studies may show either this or that the one is (much) more realistic and hence useful than the other. Hence, the paper leaves the challenge to start such comparative case studies.[21]

Department of Theoretical Philosophy
University of Groningen
Private address:
Platanenlaan 15
2061 TP Bloemendaal
Netherlands
www.rug.nl/staff/t.a.f.kuipers
T.A.F.Kuipers@rug.nl

21 I like to thank Ilkka Niiniluoto for his remarks on the manuscript.

Raffaella Campaner

Philosophy of Medicine and Model Design

Abstract

This contribution intends to show how philosophy of science and medicine have been increasingly interacting and are further called to interact in elaborating different models, in clarifying notions like explanation, mechanism and generalization in the medical context, and in tackling the relations between explaining and intervening. On these and related topics conceptual and methodological investigations on diseases and medical research and practice are both asking for philosophical reflections and questioning available philosophical notions. In what follows, I shall illustrate how philosophy of science and medicine are mutually helping and challenging each other referring to specific medical fields that have only recently become a focus for philosophy of science, such as epidemiology, psychiatry and psychiatric epidemiology.

1. Introduction

Philosophy of medicine has taken shape as a branch of philosophy of science relatively recently, with some of its most eminent first works being published roughly since the end of the Eighties. The field has been expanding in the last few years, with an increasing number of conferences, volumes and journals devoted to philosophical reflections on medicine as a science. Together with the development of philosophy of medicine as a field of enquiry, reflections have been elaborated on what the aims and topics of philosophy of medicine are supposed to be. Philosophy of science has been approaching medical research and its applications as part of a more general trend aimed at addressing *actual* scientific practice and its theoretical implications more seriously. Philosophical enquiries are also pursued due to the increasing interest of researchers and practitioners in the biomedical sciences for conceptual clarification, thus promoting interdisciplinary exchanges. Such contacts strongly stimulate philosophy of science, presenting it with new challenges and calling for further, often innovative, reflections on such topics as processes, levels of explanation, and reduction. How and to what an extent disciplinary specificities and general conceptual and methodological aspects can be "adjusted" is one of the matters on which philosophy of medicine is called to contribute.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_38,

2. Diseases as multilevel phenomena

Diseases have long been recognized as multilevel systems: not only are they brought about by a number of different factors, but such factors act at a number of different levels (e.g. the levels of cells, molecules, tissues, organs, the whole organism). After a tremendous spurt of research in molecular biology, biochemistry and genetics, recent developments in medical enquiry are encouraging a rethink of the notion of disease: such perspectives as systems biology, epigenetics, evolutionary and Darwinian medicine[1] are re-questioning the nature and aetiology of diseases, suggesting possible theoretical frameworks which can in turn affect treatments and prevention. As far as epidemiology is concerned, recent paradigms have stressed how the range of causal factors to be entertained must be broadened, and attention has focused more and more extensively on environmental and socioeconomic factors as determinants of diseases. Pathologies are believed to arise from interactions between biomolecular and genetic factors and what are deemed higher level variables: not only do viruses, pathogenic bacteria, genes, etc., bear some responsibility in causing diseases, but place of birth, education, occupation, neighbourhood of residence, income status, and many other variables are also recognized to play a prominent role. Many investigations on the socioeconomic determinants of diseases have been reported in the literature, especially in so-called "social epidemiology" and "eco-epidemiology". To mention just a few examples, studies have been performed on the causal influence of neighbourhood socioeconomic conditions on the incidence of coronary heart disease,[2] or on the relationships between area-level socioeconomic disadvantage and advanced breast cancer and cirrhosis mortality.[3]

This perspective has many implications with respect to the generally accepted view of health and disease, and also with respect to epidemiology's main goals, namely prevention, public health, and the reduction – and eventually the elimination – of health differences. Diseases are seen as multi-faceted and

1 For some reflections on similarities and differences between these two views, see Pierre-Olivier Méthot, "Research Traditions and Evolutionary Explanations in Medicine", in: *Theoretical Medicine and Bioethics* 32, 1, 2011, pp. 75-90.

2 Ana V. Diez Roux, Sharon Stein Merkin, Donna Arnett, et al., "Neighbourhood of Residence and Incidence of Coronary Heart Disease", in: *New England Journal of Medicine* 345, 2, 2001, pp. 99-106; Susan A. Everson-Rose, Tené T. Lewis, "Psychosocial Factors and Cardiovascular Disease", in: *Annual Review of Public Health* 26, 2005, pp. 469-500.

3 Peter D. Baade, Gavin Turrell, Joanne F. Aitken, "Geographic Remoteness, Area-level Socioeconomic Disadvantage and Advanced Breast Cancer. A Cross-sectional Multilevel System", in: *Journal of Epidemiology and Community Health* 65, 2011, pp. 1037-1043; Albret Dalmau-Bueno, Anna García-Altés, Marc Marí-Dell'Olmo, et al., "Trends in Socioeconomic Inequalities in Cirrhosis Mortality in an Urban Area of Southern Europe: a Multilevel Approach", in: *Journal of Epidemiology and Community Health* 64, 2010, pp. 720-727.

highly variable, and human beings are conceived of as deeply intertwined with their surrounding environment, with the society they belong to and the economic setting they live in. Public health prevention strategies are encouraged to devote much attention to improving environmental and socioeconomic conditions, and epidemiology therefore tends to intensify its exchanges with other disciplines. The promotion of this paradigm also goes hand in hand with broad reflections on the features of epidemiology as a disciplinary field and with overcoming the traditional distinction between the so-called natural and social sciences: given the aetiology of pathologies, epidemiology cannot be classified – strictly speaking – with either the natural or the social sciences, and cannot but be involved in research activities and results coming from both molecular biology and economics, sociology, and political sciences.

The acknowledgment of the role of socioeconomic factors has partly concerned psychiatry as well, and psychiatric epidemiology is also growing in importance. Socioeconomic determinants are being recognized as increasingly relevant also to mental disorders – pathologies whose origin is in most cases extremely difficult to disentangle – with significant theoretical and practical implications. On the one hand, progress in molecular biology, genetics and neuroscience has promised to yield deep insights into the workings of the human brain. Many studies have focused on the genetic-molecular components of mental disorders, neuro-chemical alterations and the role of genetic risk factors. For instance, the early-onset form of Alzheimer's disease, which affects individuals in their late forties and fifties, has been related to at least three kinds of gene mutations: mutations in APPs, the gene for amyloid precursor proteins found on chromosome 21; in the Presenilin 1 and Presenilin 2 genes, found on chromosomes 14 and 1 respectively; and in the apolipoprotein E (ApoE) gene found on chromosome 19.[4] To mention another case, studies have been performed on the relationship between the polymorphism in the COMT gene, encoding catechol-O-methyltransferase and disorders like schizophrenia and panic disorder.[5] At the same time, a causal role is increasingly attributed

4 ApoE has three allelic forms (E2, E3, E4), and it is ApoE4 in a double dose which acts to bring about the disease. The genes APP, PS1 and PS2 are regarded as deterministic causes, that is, each is sufficient in humans for the occurrence of the disease, whereas ApoE4 is regarded as a probabilistic contributory cause, increasing the likelihood of the disease in some patients. See Kenneth Schaffner, "Extrapolation from Animal Models", in: Peter Machamer, Rick Grush and Peter McLaughlin (Eds.), *Theory and Method in the Neurosciences*. Pittsburgh: University of Pittsburgh Press 2001, pp. 200-230, pp. 218-220. For a discussion of other genetic models accounting for psychiatric disorders, see Kenneth Schaffner, "Etiological Models in Psychiatry", in: Kenneth Kendler and Josef Parnas (Eds.), *Philosophical Issues in Psychiatry*. Baltimore: The Johns Hopkins Press 2008, pp. 50-90, esp. pp. 55-62.

5 Steven P. Hamilton, Susan L. Slager, Gary A. Heiman, et al., "Evidence for a Susceptibility Locus for Panic Disorder Near the Cathecol-O-Methyltranferase Gene on Chromosome 22", in: *Biological Psychiatry* 51, 7, 2002, pp. 591-601; Michael F. Egan,

to environmental and socioeconomic determinants. In the case of depression, for instance, life events involving both genetic liability and such factors as childhood parental loss, a disturbed family environment, and various stressful life events are picked out as pathogenic factors;[6] in the case of anorexia nervosa, socioeconomic roots are identified such as the high food availability and the dominant attitude towards female body size in high-income countries; in the case of schizophrenia, a causal relation has been drawn between the timing and duration of exposure to urban life and the incidence of the pathology.[7] Explanations, prevention strategies, diagnoses and treatments of mental disorders are hence urgently required to integrate results from different fields of enquiry.

3. Diseases as mechanisms

The elaboration of multilevel paradigms of both physical illness and mental disorders is being accompanied by extensive reflections on causation – questioning, among others, how causation can cross levels – and on models of explanation, prediction and control. How different level – mostly probabilistic – causes are to be conceived, how they mutually interact, how their organization can be grasped are objects of debate. Without being by any means the only one, mechanism is one of the notions most extensively analyzed to capture medical causation. Together with issues emerging in other special sciences, investigations on the aetiology and functioning of diseases have stimulated a revision of the notion of mechanism with respect to its previous uses, which mainly had physics as a key-reference discipline. No consensus has been reached so far on how exactly to define the concept of mechanism in the philosophical literature, let alone in the biomedical sciences, where undoubtedly an extremely wide, almost ubiquitous use is made of mechanistic terms. However, a mechanism can be taken roughly to be a complex system that produces a behaviour or an outcome by the interaction of a number of parts. The productive behaviour is enacted by virtue of the mutual organization of the mechanism's component parts, which must be adequately located along different

---

Terry E. Goldberg, Tonya Gscheidle, et al., "Relative Risk of Attention Deficits in Siblings of Patients with Schizophrenia", in: *American Journal of Psychiatry* 157, 2000, pp. 1309-1316. For a discussion of studies on COMT gene, see Ezra Susser, Sharon Schwartz, Alfredo Morabia and Evelyn Bromet, *Psychiatric Epidemiology*. Oxford: Oxford University Press 2006.

6 Kenneth S. Kendler, Charles O. Gardner, Carol A. Prescott, "Toward a Comprehensive Developmental Model for Major Depression", in: *American Journal of Psychiatry* 159, 2002, pp. 1133-1145.

7 Dana March, Ezra Susser, "Invited Commentary: Taking the Search for Causes of Schizophrenia to a Different Level", in: *American Journal of Epidemiology* 163, 11, 2006, pp. 979-981.

levels, properly structured and oriented, and whose activities must have a specific temporal order, rhythm and duration.

What can be regarded as the main aims, actual uses and possible advantages of employing a mechanistic approach in conceiving of pathologies? To start with, the notion of mechanism allows thinking about how disease states arise and develop to be rooted in productive webs of causes, representing multilevel structures and their organization. Mechanistic systems can also be exhibited as systems that can be decomposed according to structural or functional criteria, or both, proving very fruitful both in pursuing further biomedical research and in planning and delivering treatments. Mechanistic models can be drawn in different ways, isolating different parts of the system according to the context and purpose of investigation. Moreover, given the current definitions of mechanism, different components and activities can be admitted as constituents of mechanistic systems; bottoming out can hence be related to the specific field of medical enquiry and vary together with scientific progress and changes in medical knowledge. Given the importance attributed to the mutual organization of parts, *all* components of the mechanistic systems under investigation must be taken into account: so-called "internal" features of the organism and "external" – e.g. socioeconomic – factors cannot be investigated as separate, all being involved in generating disease. No level is regarded as more fundamental than another and all the relevant factors are considered, no matter at what level they stand, thus matching current trends in conceiving of diseases mentioned above. Identification of the component parts, their mutual organization and interactions is sometimes regarded as crucial for the very *characterization* and *classification* of pathologies. For instance, it has been argued that a thorough investigation of pathogenic mechanisms, including specific mechanistic steps in carcinogenesis, is needed to adopt an unequivocal definition of cancer.[8] In addition, the order and timing in which different causal factors act and interact, and whether myriad causes operate in parallel or in a sequence, are key-aspects in a causal analysis of diseases. An adequately devised notion of mechanism can be adopted in an attempt to grasp the *dynamicity* and *development in time* of pathologies, elucidating the *processes* going from causes to induction period, up to symptoms, and stressing relevance of the *order* in which different pathogenic factors operate. Furthermore, a mechanistic approach can be employed when patterns of *diffusion* of diseases are to be drawn: when concerned with infectious diseases, epidemiology also calls for the elaboration of models representing disease transmission, accounting for patterns of exposure and interactions within populations, for prevention and public health purposes. Investigations are hence performed regarding time-varying patterns of connections among individuals and how they affect population-level outcomes. In this respect, notions like interaction and transmission, crucial in some mechanistic views, play a very important part.

8 Paolo Vineis, Miquel Porta, "Causal Thinking, Biomarkers, and Mechanisms of Carcinogenesis", in: *Journal of Clinical Epidemiology* 49, 9, 1996, pp. 951–956.

A mechanistic approach can prove useful for various cognitive aims. Mechanisms can allow the results of studies to be generalized beyond specific tested cases, and serve to provide standard models of diseases not only for research purposes but also for diagnosis and therapy. Although they then have to be adapted to individual cases, which hardly ever mirror all aspects of the model, characterizations of the average course of the pathologies have often allowed a new taxonomy for diseases, moving on to disease definitions based on causative events and pathways: mechanistic definitions of diseases can break down "the disease heterogeneity that the phenotypic definitions had masked"[9]. On the one hand, a mechanistic representation of the pathology is provided by biomedical research as a description of its "usual" or "average" behaviour – or as the disruption of a standard healthy organism's behaviour. On the other, the individual patient is seen as a specific instantiation of such a mechanism, with many peculiarities that clinical medicine has to deal with. Mechanisms generally also have c*onfirmative power*: previously acquired mechanistic knowledge can play a confirmative role, as well as help in *hypothesizing* that some mechanistic relations hold in the investigation of still poorly understood phenomena. However, probably the most common use of mechanistic notions occurs in *explanatory* investigations, when a deeper knowledge of the phenomenon is claimed to be sought which can in turn lead to the development of better treatments and more effective preventive strategies. Medical explanations require the specification of complex mechanisms clarifying the actions and effects of pathogenic factors and exhibiting the details of the processes they trigger. When looking for a mechanistic explanation, we do not rest content with an understanding of just *what* the causal factors are, but aim to know *how* we get from the causes to the effect, i.e. what goes on *between* the causes and the effect. The claim that causal explanations have to make explicit *how* the effect has been brought about by the cause has been largely made in epidemiology, arguing, among others, that the so-called "black box", "risk factor – outcome" paradigm must be overcome, displaying what mediates from the causes to the effects and crossing boundaries between different levels of inquiry. "Epidemiologists aim not only to identify causes, but also to explain the causal processes that lead to disease". Instead, the focus on risk factor disease associations neglects "the downstream mechanisms that allow us to understand how the risk factors operates"[10]. In biomedical research a mechanism can also be traced to partly reduce the range of plausible explanations. Some alternatives are ruled out because any explanation will have to account not only for the association between the causal factors and the outcome, but also for the explanatory power of the mechanism that has been identified.

9 John I. Bell, "Clinical Research is Dead; Long Live Clinical Research", in: *Nature Medicine* 5, 5, 1999, 477-478, p. 477.

10 Susser, Schwartz, Morabia and Bromet, *ibid*., p. 416.

Possible advantages and merits of a mechanistic view notwithstanding, limits and puzzling issues are present as well, concerning, amongst other, the multilevel character of diseases. We have already insisted on how special attention must be paid to interrelationships between "internal", organismal, and "external", socioeconomic and environmental, factors, which cannot just be investigated separately. How such different levels *precisely* interact with each other is a problematic topic, having to do also with whether, e.g., biomedical research and epidemiology are committed to a *genuinely* multilevel idea of mechanism. How are higher level factors actually embodied into biochemical processes? How do they enter the etiopathogenic mechanisms in the patient? Much remains unsettled with respect to the extent to which social and economic circumstances influence disease pathways. Some studies focus, for instance, on stressful working conditions, or the perception of having a low social condition, as contributing causes of cardiovascular diseases. These factors lead to an increase in the production of cortisol and sympathetic hypertonus, and through them to an increase in the risk of myocardial infarction. Research is hence pursued on how higher level factors act by entering the biochemical processes and how they impact on individual disease states. At the same time, though, this does not have to be equated with embracing a reductionist standpoint and renouncing a genuinely multilevel paradigm of pathologies. Some non-reductionist notion of mechanism needs to be devised preserving a conception of diseases as complex systems, and accommodating interlevel integration, studied by different disciplinary fields in an attempt to elaborate as complete an explanation as possible. The challenge is to understand *how* the various factors involved interact and contribute to the global behaviour of the system. Though some theoretical progress has been made, *how* bridges are to be built across levels, and different factors integrate, still seems a largely open issue.

4. Interventionist approaches to diseases

Although mechanistic notions are widely employed, most of the times no complete mechanistic representation is reached, and all we get are mechanism sketches with gaps we do not yet know how to fill in, thereby indicating that further work needs to be done. Moreover, mechanisms are not essential for all cognitive purposes: for instance, prevention and treatment can do, and very often *have* to do, in the absence of mechanistic knowledge, or in the presence of only sketchy or very partial clues about mechanisms. Building on a manipulative notion of cause and couched in counterfactual terms, interventionist views can prove very useful in this respect. While not arguing *against* mechanicism, the interventionist approach separates issues of causation from issues to do with the underlying mechanisms. Woodward's theory, in particular, grounded on the notion of invariance under intervention and conceiving of causes as difference-makers, has been the most widely discussed in

the last decade or so, often with a specific concern for its applicability. In this view, generalizations are deemed causal if they are invariant under certain interventions; they can be more or less invariant and admit of exceptions. The generalizations we are able to frame can have a narrow scope, but as long as they are relatively stable, they have explanatory power. To explain is to answer a "what-if-things-had-been-different" question: an explanation is required to cite, ideally, all and only those factors that, if changed, would make a difference to the explanandum. Scientists' capacity to explain is thus very strictly linked to their knowledge of how to manipulate the phenomena to be explained, and to the related capacity of modifying the state of affairs. Tractable factors are hence the privileged object of causal enquiry.

What can the main advantages of a manipulative perspective be in the field at stake? A counterfactual-manipulative view appears highly suitable to account for causal claims assessed directly for practical aims, like prevention and public health. What are revealed are the factors on which we can intervene, and which, while being mechanistically relevant, often do not belong to a detailed mechanistic picture (at least, not yet). If complex systems are under enquiry, interventionist accounts see no problem in mixing causes belonging to different levels: high level variables can act on lower level ones, and vice versa, as long as stable manipulability relations hold. They can all equally figure in "what-if-things-had-been-different" explanations. Once determinants and outcomes at different levels of organization have been analyzed, decisions will be taken on what is *the most efficacious level* to act upon, to affect concrete situations. Multilevel systems can thus be admitted without worrying about what their internal workings are like, as is done, e.g., by risk-factor epidemiology. The same high level variables (e.g. one's education or occupation) might turn out to be causally involved in a number of diseases, and will thus have to be of primary concern for prevention campaigns and public health policies, but no privilege is assigned a priori to any level over the others. A crucial role will be assigned to some variables – in other terms – on pragmatic, control-oriented grounds.

Further possible merits of a manipulative-counterfactual perspective might be particularly relevant in the context of psychiatry. If Woodward himself suggests that his approach ought to be employed in psychiatry,[11] a counterfactual-manipulative account has been advocated specifically with respect to the health sciences, and especially to psychiatry, also by Kenneth Schaffner and John Campbell.[12] To start with, referring to just hypothetical interventions, this approach suits cases in

11 James Woodward, "Cause and Explanation in Psychiatry" and "Comment: Levels of Explanation and Variable Choice", in: Kendler and Parnas (Eds.)*, ibid.*, pp. 132-184 and pp. 216-235.

12 Kenneth Schaffner, "Clinical and Etiological Psychiatric Diagnoses: Do Causes Count?", in: John Sadler (Ed.), *Descriptions and Prescriptions: Values, Mental Disorders and the DSMs*. Baltimore: Johns Hopkins University Press 2002, pp. 271-290; John Campbell, "Causation in Psychiatry", in: Kendler and Parnas (Eds.)*, ibid.*, pp. 196-216.

which interventions can be very difficult or impossible to perform, for practical or ethical reasons, which is very frequently the case in psychiatry. Interventionist counterfactuals allow to identify and express those causal links we cannot *actually* intervene on. An interventionist view can also remain neutral with respect to ontological claims, and with respect to issues such as the mind-brain and mind-body problems – and all related problematic matters – which may complicate psychiatric investigations. Furthermore, in a psychiatric clinical context diagnoses are often formulated and therapies prescribed without a full understanding of the pathology's functioning. Schaffner has pointed out how diagnoses can be made in psychiatry on the basis of what variables we can control. In the case of very complex pathologies such as Alzheimer's disease, one can isolate – he points out – some causal factors by holding interfering factors constant; the dominating factor identified at any different stage will be defined as the factor exerting major influences downstream from it. "Manipulation of such a dominating factor may thus have major effects on the future course of the complex system. […] Such factors are major leverage points that can permit interventions, as well as simpler etiological explanations, which focus on such factors"[13]. While Campbell maintains that the interventionist perspective contrasts with any view on causation that takes it to be a matter of mechanistic links, such contrast does not necessarily hold; interventionist accounts can admit of mechanistic models, without committing to a mechanistic theory.

Diagnostic, treatment, prevention and public health aims can affect the search for explanations, and the choice of the most adequate, or the most likely to be achieved, account in a given context. A what-if-things-had-been-different claim can yield insights into properties we would like to control, and can hence be adopted for pragmatic purposes, when an intervention, i.e. a therapy or some prevention strategy, is to be performed. Some preference can thus be accorded, for instance, to either the closest or the most remote causal factors, depending on what we might *be able* to intervene on, or on what we might *want* to intervene on. In a sense, causes have a perspectival component. At the same time, this is not to say that research on *how* more remote, or higher level, variables enter into the pathological processes must not be pursued any further.

5. Concluding remarks: Challenges to philosophy of medicine

The cases of epidemiology and that of psychiatry show how advances in both fields go hand in hand with debates on the notion of disease, which is taken as genuinely multilevel, and its aetiology, and with enquiries on how different factors become embodied and/or on how they can be controlled. Investigations from, among

13 Schaffner, "Clinical and Etiological Psychiatric Diagnoses: Do Causes Count?", *loc. cit*., p. 286.

others, physiology and cancer research are also stimulating some re-thinking of philosophically relevant medical issues. In physiology some emphasis is being put, for instance, on the need to overcome a gene-centered perspective, adopting an integrative approach to networks of interactions between genes and higher level functions of organisms. The integration of molecular genetic information and higher level components is suggested as a possible route to assess a multi-level viewpoint in physiology, and to reconnect physiology with developmental and evolutionary biology.[14] With respect to cancer research, some of the latest interpretative models stress how different levels of biological organization are involved in the aetiopathogenesis of tumours – complex and heterogeneous diseases – and in their progression, with major implications not only for investigation in molecular biology, but also, for instance, from an epidemiological and clinical standpoint. In both the physiology and cancer research cases, what is pointed to is the crucial role of an adequate comprehension of different levels of causal explanation, and traditional reductionist stances are challenged, looking at the organism as a whole. At the same time, "ontological and epistemological motivations to continue the reductionist program" are not completely bracketed: what is required are "novel ways to think about networks where non-linearity becomes the rule of biological processes and functions, not the exception" [15].

Reflections from various medical disciplines are hence questioning the conceptualization of diseases and can challenge some important notions in the philosophy of science, contributing to their revision. With respect to the notion of mechanism, a largely adaptable notion is to be elaborated, with possible contributions from different neo-mechanistic positions, both quite flexible and capable of accounting for disciplinary specificities. With respect to generalizations describing system behaviours, they are presented as narrow in scope, often quite specific to a given system, stochastic, contingent and admitting of exceptions. With respect to biomedical theories, they can be defined as "middle-range theories" [16], that is, "overlapping interlevel models". They are supported by mutually reinforcing types of evidence, aim to explain events at different levels and to create bridges among such levels, refer to processes developing in time, have a limited scope and admit of variations. While restricted, they play a fundamental role, for instance with respect to the adoption of animal models, to draw analogies between different organisms, to extend results of investigations to new cases, and for experimental, predictive and control purposes.

A focus on both the specific medical field one is working in and the purpose of the enquiry has also led to pluralistic approaches to some of the most contro-

14 Denis Noble, "Neo-Darwinism, the Modern Synthesis and Selfish Genes: Are They of Use in Physiology?", in: *Journal of Physiology* 589, 5, 2011, pp. 1007-1015.

15 Marta Bertolaso, "Towards an Integrated View of the Neoplastic Phenomena in Cancer Research", in: *History and Philosophy of the Life Sciences* 31, 2009, pp. 79-98, p. 90.

16 Kenneth Schaffner, *Discovery and Explanation in Biology and Medicine*. Chicago: The University of Chicago Press 1993.

versial points, overcoming long-established philosophical categories. With respect to the reductionist-antireductionist issue, for instance, it has been suggested that "'mixed' models are likely to be the most useful for the foreseeable future, thus obviating a precise definition of reductive versus nonreductive"[17]. With respect to explanation, explanatory pluralism seems preferable to monistic explanatory approaches to account for the complexity of systems under enquiry and for the multiple targets for which an explanation can be sought. Explanatory pluralism admits multiple mutually informative perspectives, focusing on different levels of organization and abstraction, and promotes different and complementary kinds of understanding of different kinds of pathologies. Adopting an explanatory pluralistic stance is accompanied by much emphasis on the context, which is held to significantly condition the choice of the best suited approaches. The object of a medical investigation (i.e. a "standard" organism, the single patient, a given population, a given subset of the population, …) and its purpose (e.g. whether we are aiming to explain mechanistically, or to intervene to predict and control, …) must be considered, and the choice of a given model will be driven both by what it is a model *of* and by what it is a model *for*.

The encounter with medicine asks philosophy of science to address new challenges (e.g. to deal with very complex definitions of health and disease, with articulate interactions among different levels, and with interfield theories); to re-think traditional philosophical problems and test old notions with respect to new frameworks and specific instances (e.g. the reductionism/anti-reductionism issue; the relationship between explaining and intervening); to contribute more significantly to the clarification of specific medical problems (e.g. the elaboration of a logic of diagnosis; the relation between general biomedical models and clinical individual variations). If descriptions, explanations, diagnoses and therapeutic strategies can be in many respects partial and "patchy", philosophical analysis can contribute to conceptual refinement and proper organization and integration of the "patchy bits". As Kenneth Kendler suggests referring to psychiatry, investigations should have as a goal "'piecemeal integration' in training to explain complex aetiological pathways to illness bit by bit"[18].

The adoption of certain descriptive, explanatory and predictive models and given sets of notions is accompanied by the shaping of different models of health and disease, which in turn can promote the development of certain trends in biomedical research, and/or enhance different aspects of medical practice and public health policy, and thus have a strong impact on both medicine as a science and our lives as patients. The adoption of pluralistic conceptions does not confine us to a single level of analysis, and hence does not confine us to any single level of action. Accounting for disciplinary specificities and aims, some comprehensive frame-

17 Schaffner, "Etiological Models in Psychiatry", *loc. cit*., p. 51.

18 Kenneth S. Kendler, "Toward a Philosophical Structure for Psychiatry", in: *American Journal for Psychiatry* 162, 2005, pp. 433-440, p. 433.

work must be provided that allows for a flexible conceptualization of disorders, able to absorb further discoveries on their constitution and aetiology, and trying to meet the challenge of keeping together the visions of biomolecular research and of health care.

Acknowledgments: I would like to thank Marta Bertolaso and Matteo Cerri for their helpful input.

Department of Philosophy
University of Bologna
Via Zamboni 38
40126, Bologna
Italy
raffaella.campaner@unibo.it

Roman Frigg, Seamus Bradley, Reason L. Machete
and Leonard A. Smith

Probabilistic Forecasting: Why Model Imperfection Is a Poison Pill

Abstract

Foretelling the future is an age-old human desire. Among the methods to pursue this goal mathematical modelling has gained prominence. Many mathematical models promise to make probabilistic forecasts. This raises the question of exactly what these models deliver: can they provide the results as advertised? The aim of this paper is to urge some caution. Using the example of the logistic map, we argue that if a model is non-linear and if there is only the slightest model imperfection, then treating model outputs as decision relevant probabilistic forecasts can be seriously misleading. This casts doubt on the trustworthiness of model results. This is nothing short of a methodological disaster: probabilistic forecasts are used in many places all the time and the realisation that probabilistic forecasts cannot be trusted pulls the rug from underneath many modelling endeavours.

1. Introduction

Foretelling the future is an age-old human desire, and the methods to pursue this goal are varied. Ancient Greeks consult an oracle; the superstitious ask a fortune teller to read the cards, and the rationally minded revert to scientific methods. Among the methods of science, mathematical modelling has gained prominence: from planetary motion to nuclear fission; and from the growth of a population to the returns of an investment, there is hardly a phenomenon that has not at one point or other been modelled mathematically. Many of these models make probabilistic forecasts: they provide us with probabilities for certain future events to occur. Weather models, climate models, financial market models, and hydrological models are but some prominent examples of models making probabilistic predictions. Designing such models is aided by the availability of ever increasing computational power, which has led to a trend of building ever larger and more complex models which are capable of making ever more precise predictions on an ever finer scale.

An example of the use of such a model is the recent project called *United Kingdom Climate Projections* (UKCP), which aims to make high resolution

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_39,

probability forecasts of the climate for up to 2100. Figure 1 provides an example of such a forecast. It shows probabilities for different changes in precipitation under a medium level emission scenario.[1] The figure tells us, for instance, that there is a 0.5 probability for a 20–30% reduction in precipitation in London by 2080. One of the striking aspects of this prediction is its precision. Calculations are made for a high resolution grid and so the forecast is able to distinguish, for instance, between the effects of climate change in London and Oxford (which are only an hour apart by train).

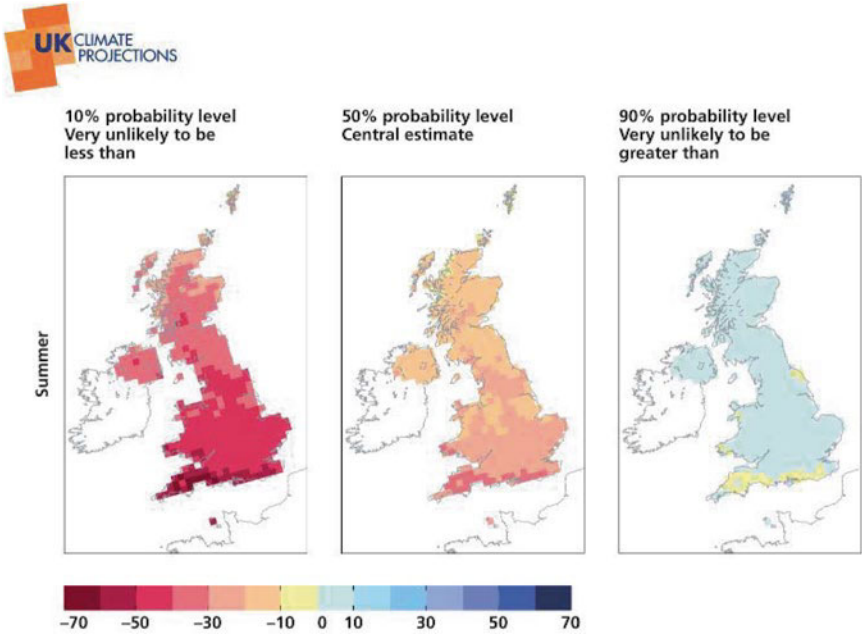


Figure 1. Change in summer mean precipitation (%) for the 2080s under a medium emission scenario. Source: UKCP.[2]

Computational power does not come for free. Super-computers are expensive tools, and developing and operating large computational models takes up the best part of the working hours of an ever increasing number of scientists. This raises the question of exactly what these models deliver: can these models provide the results as advertised?

1 UKCP uses the IPCC A1B scenario. This is a kind of "optimistic" scenario of rapid growth and then a levelling off of the population by 2050 and a balance of renewable and fossil fuel energy. Total cumulative emissions amount to roughly twice what cumulative emissions were in 1990.

2 http://www.ukcip.org.uk/wordpress/wp-content/UKCP09/Summ_Pmean_med_2080s.png; retrieved on 12 October 2011.

The aim of this paper is to urge some caution. We argue that if a model is non-linear and if there is only the slightest model imperfection, then treating model outputs as decision relevant probabilistic forecasts can be seriously misleading. This casts doubt on the trustworthiness of model results like the one we have just seen. In what follows we discuss this claim with a focus on climate modelling; we do so for the purpose of illustration and emphasise that the problem we describe crops up in all phenomena best modelled by non-linear models.

We begin by outlining the general methodology used in producing probabilistic forecasts, which we refer to as the *default position* (Section 2). Using computer simulations in a simple model we show that the default position produces seriously misleading results if the dynamics of the system is non-linear (Section 3). This casts serious doubt on the trustworthiness of model-based probability distributions, and there is unfortunately no quick and easy way to dispel these doubts (Section 4). This raises serious questions about how models are (and indeed should be) used to make informed policy decisions (Section 5).

2. The default position

From a formal point of view, a climate model is a dynamical system, which we denote by (X, φ_t, μ). As the notation indicates, a dynamical system consists of three elements. The first element, X, is the system's *state space*, which contains all states in which the system could be. What these are depends on the nature of the system. For instance, the state space of a particle moving along a straight line are the real numbers, and the one of a hockey puck sliding on a square ice rink the unit square. The second element, φ_t, is the *flow* (or *time evolution*): if the system is in state $x \in X$ now, then it is in $y = \varphi_t(x)$ at any later time t. In other words, φ_t, tells us how the system's state changes over the course of time. The third element, μ, is the system's Lebesgue measure: it allows us to say that parts of X have a certain size. In case X is the real axis, μ is the length of an interval, and if X is the unit square the measure informs us what the area of parts of the square are. In the argument to follow μ plays no role and we note it here merely for the sake of completeness.

In the case of climate models X consists of relevant weather variables (such as air temperature, precipitation, wind speed, …), and φ_t tell us how they change over time. When descibed at that level of abstraction, one could be left under the impression that climate models are rather simple things. It is important to counter this impression before it gains traction. A full specification of the system's state space would involve giving the air temperature, precipitation, etc. at *every point* in the atmosphere of the earth! It is not only a practical impossibility to obtain these data; it is also an impossibility to store them with digital technology. For this reason we discretise the state space, meaning that we put a grid with a fininte number of cells on X and represent the state of an *entire cell* by one set of values for the relevant

variables. The grid size is the length of the sides of the cells. Typically the grid size used in a climate model is well over 100km. Covering the world with such a grid still leaves us an enormous amount of data! Yet it is important to emphasise that the volume of numbers notwithstanding, this is a rather coarse description. For instance, the weather in the entire city of London is now represented by one set of numbers (one number for temperature, one for precipitation, etc.). The dynamics of the model raises even more issues. In order to specify φ_t we have to make a number of strongly idealising assumptions: we distort important aspects of the topography of the surface of the earth as the resolution of these models does not allow for realistic mountain ranges like the Andes, does not resolve the southern half of the state of Florida, many islands simply don't exist, including small volcanic island chains easily visible in satelite photographs due to their interaction with clouds, and of course cloud fields themselves are not modelled realistically. Based on these idealising assumptions we can use basic physics (essentially fluid dynamics and thermodynamics) to formulate the equations of motion for the simplified earth's climate system. These equations are non-linear and we cannot solve them analytically. For this reason we resort to the most powerful computers available to compute solutions. The result of these computer simulations is φ_t.[3]

The formal apparatus developed so far has it that the flow takes as input a particular initial state x and then tells us into what state $y = \varphi_t(x)$ this condition evolves under the dynamics of the system. Unfortunately this algorithm is not very useful in practice because we never know in what *exact* state the system is (if such a thing exists at all). To begin with, there is no measurement device that provides exactly correct values and so every measurement result comes with a certain margin of error. But more importantly, there is no such thing as *the* true wind speed in a model grid point corresponding to central London! All we can truthfully say is something like 'the wind speed at a particular random location within that grid cell is likely to lie within a certain range'. We account for some of these uncertainties by specifying a probability distribution $p_0(x)$ over initial states, where the subscript indicates that the distribution describes our uncertainty about the initial condition at $t = 0$. There is of course a legitimate question about what the correct distribution is; we set this issue aside and assume that in one way or another we can come by the correct $p_0(x)$ (in the sense that it is a correct representation of our uncertainty).[4]

---

3 For a general introduction to climate modelling see Kendal McGuffie and Ann Henderson-Sellers, A Climate Modelling Primer. 3rd ed. New Jersey: Wiley 2005; a discussion of the specific models used in UKCP can be found at http://ukcp09.defra.gov.uk/.

4 For a discussion of different kinds of uncertainty and their sources see Seamus Bradley, "Scientific Uncertainty: A User's Guide", in: Grantham Institute on Climate Change Discussion Paper 56, 2011 (available at http://www2.lse.ac.uk/GranthamInstitute/publications/WorkingPapers/Abstracts/50-59/scientific-uncertainty-users-guide.aspx), and Leonard A. Smith and Nicholas Stern, "Uncertainty in Science and its Role in Climate Policy" Philosophical Transactions of the Royal Society A 369, 2011, pp. 1-24.

The question then becomes: how does $p_0(x)$ change over the course of time? The flow φ_t can now be used to move $p_0(x)$ forward in time: $p_t(x) := \varphi_t[p_0(x)]$.[5] This distribution is the central item of the *default position*, the view that we obtain the decision-relevant probabilities for certain events to occur by plugging the initial distribution into the model and using the flow to obtain forecast probabilities for events at later times. The qualification 'decision-relevant' is crucial. The default position does not make the (trivial) statement that $p_t(x)$ is a probability distribution in a formal sense (i.e. that it is a mathematical object satisfying the axioms of probability); it is committed to the (non-trivial) claim that these probabilities are the correct probabilities for outcomes in the world in the sense that a rational decision maker should adjust his/her beliefs to these probabilities and act accordingly (assuming that there is no other pertinent evidence). In other words, $p_t(x)$ is taken to provide us with predictions about the future of sufficient quality that we ought to place bets, set insurance policies, or make public policy decisions according to the probablities given to us by $p_t(x)$.[6]

3. The poison pill

Its intuitive appeal notwithstanding, the default position is wrong: $p_t(x)$ need not be the correct probability distribution, and taking $p_t(x)$ as a guide to actions can be ruinous. Our strategy is to present a case where one can explicitly see that $p_t(x)$ need not be the correct probability distribution. This is enough to refute the default position, which has it that $p_t(x)$ *always* is the correct probability distribution.

Consider the following thought experiment. McMath has a pond in his garden where he breeds fish. He does not like being a hostage to fortune and wants to plan carefully how much food he will have to buy to feed his fish. To this end he constructs a model which allows him to predict the size of the population in his pond at a given time. He first introduces the population ratio ρ_t: the number of fish in the pond at time t divided by the maximum number of fish the pond could accommodate; ρ_t lies in the unit interval [0, 1]. To predict future populations he comes up with

5 We use square brackets to indicate that $\varphi_t[p_0(x)]$ is the propagating forward in time of the initial distribution $p_0(x)$. The flow of distribution derives from the flow of a state as follows: $p_t(x) := \varphi_t[p_0(x)] = \sum_i p_0(z_i)$, where the sum of z_i reflects each of the states in X which are mapped onto x under the flow φ_t (i.e. $\varphi_t(z_i) = x$ for all i); if the flow is invertible this reduces to $p_t(x) = p_0(\varphi_{-t}(x))$.

6 UKCP09 probabilities are formed in a more complicated manner, combining outputs from multiple (imperfect) models using Bayesian methods (see http://ukclimateprojections.defra.gov.uk/23239 and http://ukclimateprojections.defra.gov.uk/23210). However, it is unclear why combining the outputs of several structurally imperfect models should make the problems we describe in the following section go away.

$$\rho_{t+1} = 4\rho_t(1-\rho_t), \tag{1}$$

where time t is measured in units of weeks. So the model says that the population ratio in a week's time is four times today's ratio multiplied by one minus today's ratio.[7]

This allows McMath to predict the future size of his population given he knows today's size. The model is a dynamical system in the above sense with the unit interval [0, 1] being the state space, the flow being given by Equation 1, and the measure being the "usual" length of real intervals. So McMath decides to follow the prescription of the default position: he puts a probability distribution $p_0(x)$ over the initial conditions – here today's population ratio – and moves it forward in time under the dynamics of the system. He then uses the predictions thus generated to bet with one of his fellow villagers. The bet is "above or below 0.5": they split the unit interval into two equal parts, (0, 1/2) and (1/2, 1), which they call A and B respectively, and bet on whether A or B occurs in two months' time.

How successful will McMath be? Will he feed his fish well and will he win the bet against his mate? At this point the second part of our thought experiment begins: as we are pondering this question, we are incredibly lucky: heaven opens and God whispers the formula of the world's *true* dynamics into our ear:

$$\tilde{\rho}_{t+1} = 4\tilde{\rho}_t(1-\tilde{\rho}_t)\left[(1-\varepsilon)+\tfrac{4}{5}\,\varepsilon(\tilde{\rho}_t^2-\tilde{\rho}_t+1)\right] \tag{2}$$

where ε is a parameter taken here to be 0.1. We immediately realise that this is just McMath's model plus a small perturbation. Figure 2 shows both the model (Equation 1) and the world (Equation 2), which makes obvious how similar the two are.

7 Equation (1) is of course just the well-known logistic map. The rationale for choosing this equation is that it is one of the simplest non-linear maps and that it has originally been proposed as a population model; see Robert May, "A Simple Mathematical Equation with very Complicated Dynamics", in: Nature 261, 1976, pp. 459-469. For the ease of presentation we assume that a new generation of fish is born once a week.

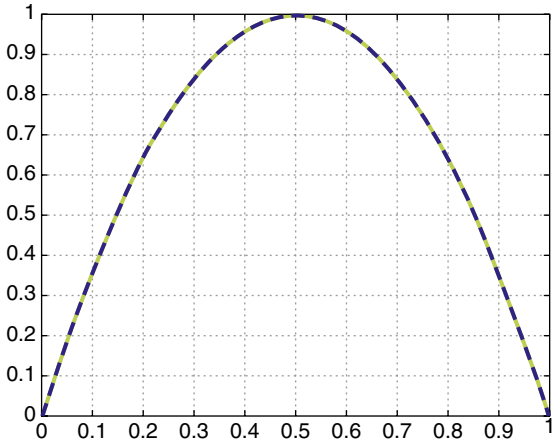


Figure 2. Equation 1 in blue (dotted line) and Equation 2 in yellow (drawn line) with ρ_t and $\widetilde{\rho}_t$ on the x-axis and ρ_{t+1} and $\widetilde{\rho}_{t+1}$ on the y-axis.

The maximum error of the model is 5×10^{-3} at $x = 0.85344$, where $\rho_{t+1} = 0.50031$ and $\widetilde{\rho}_{t+1} = 0.49531$. This error is really small, which would lead us to believe that McMath's predictions should be accurate, and that therefore the use of the default position should be a winning strategy.

But calculations are better than intuitions, and so we use our God-given insider knowledge to see how well McMath will do. We move the initial distribution $p_0(x)$ forward in time both under the dynamics of the model (Equation 1) and the world (Equation 2), which gives us the two distributions $p_t^m(x)$ and $p_t^w(x)$ for the model and the world respectivley. If the default position was correct, one would have to find that $p_t^m(x)$ and $p_t^w(x)$ are identical, or at least broadly overlap. This is because, by assumption, the initial distribution is the correct distribution and the dynamics of the world is the true dynamics, hence $p_t^w(x)$ is the correct distribution and $p_t^m(x)$ captures what happens in the world only to the extent that it agrees with $p_t^w(x)$.

Since we don't know how to calculate $p_t^m(x)$ and $p_t^w(x)$ with pencil and paper, we resort to computer simulation. To this end, we divide the system's state space into 49 cells (which, in this context, are usually referred to as 'bins'). We then choose an initial distribution of 1024 points which is distributed according to the invariant measure within a radius of 7×10^{-3} from the true initial condition. The true initial condition is randomly chosen; in the concrete example to follow it happens to lie in the third bin. The true initial condition was within the same interval, but not necessarily at the centre. In turn we iterated forward in time 1024 points from the initial distribution (see first graph in Figure 3) under each of dynamical laws. The other graphs in Figure 3 show how many points there are in each bin after two, fours and eight weeks respectively. Dividing these numbers by 1024 yields an estimate of the probability for the system's state to be in a particular bin.

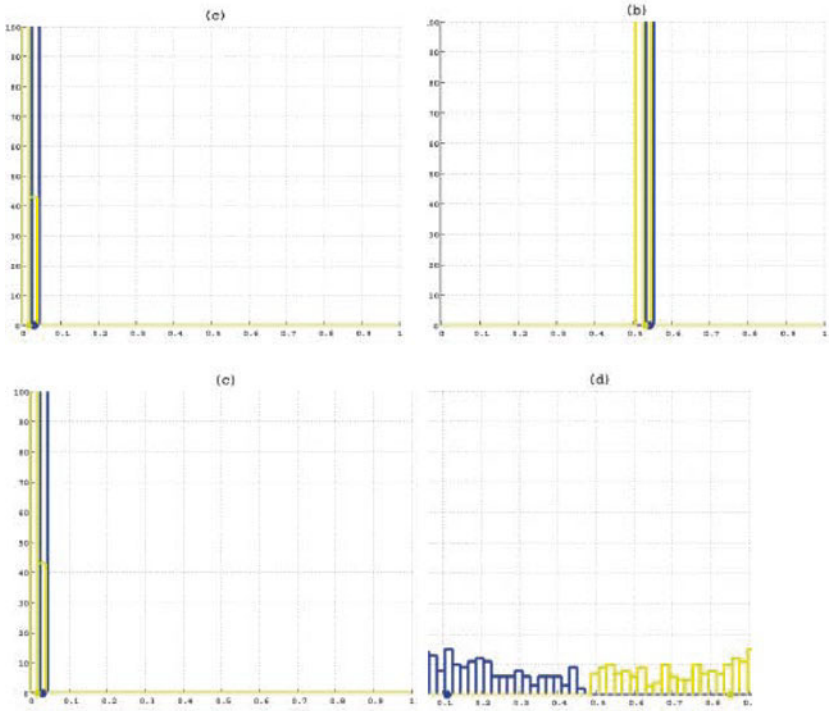


Figure 3. The evolution of the initial probability distribution under the dynamics of the model (Equation 1, blue line) and the world (Equation 2, yellow line).

These calculations show the failure of the default position. While the two distributions overlap relatively well after two and four weeks, they are almost completely disjoint after two months. The implications of this for McMath are dramatic. His calculation led him to believe that after two months $p(A) \cong 1$ and $p(B) \cong 0$ (this is read off from the blue line in the fourth graph in Figure 3). This led him to offer extremely long odds on A.[8]

But the correct probabilities (read off from the yellow line in the same figure) are $p(A) \cong 0.1$ and $p(B) \cong 0.9$. So he is very likely to lose a large amount of money to his fellow villager![9]

The moral of this thought experiment is that if a non-linear model differs from the truth only by a little bit (i.e. if the model has only a slight structural imperfec-

8 We use so-called odds-for, which give the ratio of payout to stake. These are convenient because they are reciprocals of probabilities; i.e. if $p(A)$ is the probability of A, then $o(A) = 1/p(A)$ are the odds on A.

9 Notice that our argument does not trade on worries about $p_0(x)$. We assume that the initial distribution gives us the correct initial probabilities and that setting ones degrees of belief in accordance with these probabilities would be rational. The core of our concern is what happens with these probabilities under the time evolution of the system.

tion), then probabilistic predictions can break down. This implies that the default position is wrong. Simply moving forward in time an initial distribution under the dynamics of the model will not yield decision-relevant probabilities! But the break-down of the default position is nothing short of a methodological disaster: as we mentioned above, the default position is used in many places all the time and the realisation that probabilistic forecasts cannot be trusted pulls the rug from underneath many modelling endeavours. One can sum up the result of our story in the slogan that model imperfection is a poison pill.

An immediate reply would point out that we have biased the presentation of the case in various ways to arrive at our conclusion and that the situation is in fact less dire than we make it out to be. The first bias is the focus on the two months forecast: had we focussed on the one month forecast McMath's forecasts would have been accurate enough to make both his planning and betting sustainable. Perhaps, perhaps not. In the real world heaven doesn't open and no one whispers true dynamical laws into our ears. So we cannot simply compare the model with the true dynamics and affirm that we are fine at $t = 4$. In fact, if we knew the true dynamics we would not need a model in the first place! All we have is a model, and we know that the model is imperfect in various ways. What the above scenario shows is that model-probabilities and probabilities in the world can come unstuck dramatically, and as long as we have no means of telling *when* this happens, we better be on guard! For all we know there is no method of predicting when the model is accurate other than knowing the truth in advance, in which case we would not bother with a model anyway.

The second alleged bias is the choice of the initial distribution. In order to run the calculations we have to choose a particular initial distribution (1024 points distributed according to the invariant measure within a radius of 7×10^{-3} from the true initial condition, which came to lie in the third bin). However, so the argument goes, this must be a special case that we have carefully chosen in order to drive home our sceptical conclusion; most of the distributions do not behave in this way and models provide trustworthy results most of the time. Our story, so the counter continues, shows at most that every now and then unexpected results happen, but does not warrant a wholesale rejection of the default position.

There is no denying that our calculations rely on a particular initial distribution, but that realisation does not rehabilitate the default position. We repeated that same calculation with a large number of randomly chosen initial distributions, and it turned out that about one third of these distributions showed behaviour similar to that seen in Figure 3; another third resulted in forecast distributions that manifested an overlap of about fifty percent; and only one third behaved roughly as the default position would lead us to expect. Hence, the problems we describe are by no means as rare as those critics would have it, and as long as we have no systematic way of drawing a line between the good and the bad distributions, we had better not rely too heavily on our calculations when making provisions for the future.

Some may have started wondering what all this has to do with modelling in the sciences; after all, what we care about is the future climate or the stability of financial markets and not the fishing success of an imaginary Scotsman. Unfortunately the connection between our imaginary scenario and 'real' scientific cases is tighter and more immediate than we would like it to be. The problems arise if models are non-linear and imperfect, and many scientific models have these properties. Without question, climate models have both these properties. It is not clear how to interpret the situation when different models agree (give indistinguishable probability forecasts), but in the climate case the different models give very different distributions (cf. the last IPCC Report, WG I) and so we know that the details of the models have a significant impact on expected results.[10] So when calculating, say, monthly precipitation in the 2080s based on climate models we may well not fare better with our planning of flood provision and water systems than McMath with bets.

4. Antidote wanted

The first serious issue is whether Equations 1 and 2 are good proxies for all other non-linear systems. Equation 1 is of course the well-known logistic map with the independent parameter set equal to 4, which results in the dynamics being chaotic;[11] Equation 2 is a perturbed version of it. By saying that climate or finance models will face the same predictive breakdowns we implicitly assume that the problems when making predictions with the logistic map are typical of non-linear models and will also occur in systems with completely different dynamical laws (as long as they are non-linear). It is fair to say that there is no hard and fast argument for this conclusion. However, it seems to us that the burden of proof lies with those who want to argue that the default position does not run into the problems we describe when used in the context of other non-linear models. Since the rise of chaos in the 1980s a bewildering array of non-linear systems has been studied and the general moral to be drawn from these studies is that random properties of systems get more dominant as (a) parameter values controlling the non-linear terms increase and (b) the size of the systems increases.[12] Generalising from these cases

10 See Leonard Smith, "What Might We Learn from Climate Forecasts?", in: *Proceedings of the National Academy of Science* USA 4, 99, 2002, pp. 2487-2492.

11 See Robert May, "A Simple Mathematical Equation with very Complicated Dynamics", *loc. cit.* and Leonard Smith, *Chaos. A Very Short Introduction*. Oxford: Oxford University Press 2007.

12 By 'random properties' we mean, for instance, properties belonging to the ergodic hierarchy such as being mixing or Bernoulli; for a discussion of these see Joseph Berkovitz, Roman Frigg, and Fred Kronz, "The Ergodic Hierarchy, Randomness and Chaos", in: *Studies in History and Philosophy of Modern Physics* 37, 2006, pp. 661-691. An example of a system that becomes increasingly random as the perturbation pa-

one would expect that climate models, which are both strongly non-linear and huge, should display more rather than less of the problems we have seen above.

Another set of issues concerns lead times. Three challenges can be mounted. The first points out that all we are interested in are short term predictions and the above results show that in the short term the model forecasts appear accurate – hence there is no cause for concern. In some cases this seems to be the right response. In weather forecasting, for instance, we are mainly interested in predicting the immediate future and hence limiting model runs to the short term is the right thing to do. But this response does not work in all cases. In both weather and climate modelling, for instance, we also are interested in the medium or long term behaviour and so we cannot limit predictions to short lead times. Of course what counts as short-term or long-term is relative to the model and it could be the case that by the standards of the relevant climate models a prediction for 2080 is still a short term prediction. We are doubtful that this is the case. Indeed, it would be surprising if such predictions would turn out to be short term by the lights of a model used to make that prediction, in particular given that state of the art climate models differ even in terms of their performance over the past century. Again the burden of proof lies squarely with those who believe that this is the case.

The second challenge argues for the opposite conclusion: what we are interested in is long term behaviour and so we can actually do away with detailed predictions completely and just study the invariant measure of the dynamics because it is the invariant measure that reflects a system's long term behaviour. Implicit in this proposal is the assumption that the invariant measures of similar dynamical laws are similar, because unless Equations 1 and 2 have similar invariant measures there is no reason to assume that adjusting beliefs according to the invariant measure is less misleading than adjusting them according to $p_t^m(x)$. However, it is at best unclear whether this is so. There is no proof that invariant measures have this property. Nonlinear systems are not expected to be structurally stable in general, which suggests that invariant measures need not be similar. And what is worse still, unlike McMath's pond, the world's climate is a transient system and as such it does not have an invariant measure at all.

The third challenge is that we are playing fast and loose with the notion of prediction. While McMath wants to predict what happens exactly two months from now, the above climate prediction is an average for the 2080s. So we would be comparing apples and pears. Not quite. What UKCP provides are not decadal forecast distributions. They provide an average for every year (the claim being that the distribution is the same in each year of the 2080s). This is not so different from weekly predictions in the fish model. Other predictions made by UKCP include

---

rameter is turned up is the Hénon-Heiles system; see John Argyris, Gunter Faust, and Maria Haase, *An Exploration of Chaos*. Amsterdam: Elsevier 1994. For a discussion of systems that become more random as the number of particles increases see Roman Frigg and Charlotte Werndl, "Explaining Thermodynamic-Like Behaviour in Terms of Epsion-Ergodicity", in: *Philosophy of Science* 78, 3, 2011, pp. 628–652.

the forecasts for the hottest day in August of a particular year. So what UKCP provides are not long term averages and hence an appeal to averages does not help circumventing the difficulties we describe.

5. Conclusion

We have argued that the combination of non-linear dynamics and model imperfection is a poison pill in that it shows that treating model outputs as probabilistic predictions can be seriously misleading. Probabilistic forecasts are therefore unreliable and do not provide a good guide for action as such.

This raises two questions. The first concerns the premises of the argument. The model being non-linear has been an essential ingredient of our story. While this assumption is realistic in that many relevant models have this property, there is still a question whether the effects we describe are limited to non-linear models. Arguably, if the world was governed by linear equations, then imperfect linear models need not suffer from the effects we discuss so badly. One might like to avoid the assumption that the world is governed by any equations, of course, but the relevant point here is the role of model imperfections: a linear model will suffer from these effects unless its imperfections are also linear. The model being linear does not remove the difficulty we note. And of course, in practice the best models are rarely linear, nor are the relevant laws of physics.

The second question is what conclusion we are to draw from the insight into the unreliability of model-based probabilities. An extreme reaction would be to simply get rid of them. But this would probably amount to throwing out the baby with the bathwater because, as we have seen, in about one third of the cases the model indicates usefully. So the challenge is to find a way to use the model when it provides insight while guarding against damage when it does not. Finding a way of doing this is a challenge for future research.

Acknowledgments: We would like to thank audiences in Athens, Bristol, Ghent, London, Paris, and Toronto for valuable discussions. This research was supported by the Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re. Smith would also like to acknowledge support from Pembroke College Oxford. Frigg would like to acknowledge support from the the Spanish Ministry of Science and Innovation through the project FFI2008-01580. Machete had financial support from RCUK Digital Economy Programme via EPSRC grant EP/G065802/1.

Roman Frigg
Department of Philosophy
London School of Economics
Houghton Street
WC2A 2AE, London
UK
r.p.frigg@lse.ac.uk

Reason L. Machete
Department of Mathematics and Statistics
University of Reading, P.O. Box 220
RG6 6AX, Reading
UK
r.l.machete@reading.ac.uk

Seamus Bradley
Department of Philosophy
London School of Economics
Houghton Street
WC2A 2AE, London
UK
s.c.bradley@lse.ac.uk

Leonard A. Smith
Centre for the Analysis of Time Series
London School of Economics
Houghton Street
WC2A 2AE, London
UK
l.smith@lse.ac.uk

Daniel Andler

Dissensus in Science as a Fact and as a Norm

Abstract

Dissensus – incompatible theories co-existing for an extended period – has been traditionally viewed as a rare accident or else as a stage in the progression of scientific inquiry that is bound to terminate: on these views, consensus is the stable state to which science tends. Following Miriam Solomon's reconsideration of dissensus as rationally on par with consensus, it is argued that the persistence of dissensus is compatible with the pull towards the resolution of inconsistency. While the social turn in philosophy of science goes some way towards relieving the tension, the key move is to go one step beyond and to distinguish between the social-psychological level, where the pull towards resolution is in force, and the public level, where it does not operate directly and can be counter-balanced by other mechanisms. An added benefit of this approach is to provide a more realistic picture of the scientists' predicament, at both the individual and communal levels, who face not only Nature but public science that stand in need of interpretation. Finally, it is suggested that dissensus enhances the ability of public science to quickly overcome impasses.

Science tracks truth: it aims at knowing, for any proposition of interest P, whether P is true or whether not-P is. As they cannot be both true, science cannot accept for any length of time a situation where some scientists hold P and others non-P. In other words, science cannot countenance dissensus, and if and when dissensus in fact arises, science seeks to eliminate it by putting more effort into finding out which of P or not-P is true. In fancier terms, "consensus is the *telos* of science" (Alan Richardson[1]). In fact, or so Richardson argues (in Peirce's and his own name), inquiry would not *make sense* if consensus were not the *goal* of science. Without going that far, we are at least strongly inclined to think that inquiry being constitutively the search for truth, and truth being indivisible, inquiry necessarily leads either, in case of failure, to ignorance or error, or, in case of success, to consensus.

Against this view, arguably the majority view in philosophy of science, Miriam Solomon and a few others before her have argued that dissensus is not the

1 Alan Richardson, "Solomon's Science without Conscience, or, On the Coherence of Epistemic Newtonianism", in: *Perspectives on Science* 16, 2008, pp. 246-252.

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2_40,

delendum of science[2]: although, all things being equal, consensus might be preferable to dissensus, this preference should not trump other norms; in fact, dissensus may well be rationally preferable to consensus on some occasions, and not merely for instrumental reasons, as a means to facilitate the realization of science's ultimate goal.

Now there is a way of accommodating the view of science as truth-tracking and consensus as a desirable (or ultimately inevitable) state of affairs with the view that dissensus is acceptable (or inevitable) in science. It involves tampering with the concept of truth, making it relative, partial, perspectival etc. (so that P and not-P can both be true). This – which I'll label 'relativism'– is a path that I will not follow in this paper. Instead, I aim to show that of the three posits:

1. Science aims at truth.
2. Dissensus is normatively acceptable.
3. Relativism is false.

none needs to be discharged.

I will rely on two moves to establish the consistency of the three: (1) I will distinguish between subjective and public scientific knowledge (in a sense which I will explain); (2) I will attempt to discredit the picture picture of science. On the way, I will demolish another picture, in which the scientific process is seen as a confrontation between the scientist, alone or as part of a team, and Nature. Finally, I will briefly suggest a reason for thinking of dissensus not just as a fact of life (be it a normative fact of life), but so to speak as a guardian angel of science.

1. The current views of dissensus

Dissensus (as I will be using the word) refers to a state of scientific knowledge where two (or more) incompatible theories co-exist. Dissensus implies non-agreement of course, but not any case of non-agreement counts as dissensus, which requires a certain depth and a certain permanence; dissensus does not necessarily go together with criticism, which requires acting argumentatively against a different view; let alone with controversy, which involves two camps combating one another for an extended period; or even with dissent, if dissent is understood as focusing on a view different from one's own and rejecting it.

There are three main accounts of dissensus which are currently visible, although the first and second tend to recede under the pressure of the more recent third.

2 Miriam Solomon, *Social Empiricism*. Cambridge (Mass.): The MIT Press 2001; Miriam Solomon, "Norms of Dissent", in: Damien Fennell (Ed.), *Contingency and Dissent in Science Project Discussion Paper Series*. Technical Report 0908, CPNSS, LSE, 2008.

i. Dissensus as short-lived accidents

As Marcello Pera writes at the beginning of his contribution to an important, fairly recent volume on scientific controversies, "the Founding Fathers[3] were deeply attached to the idea that science is uncontroversial"[4]. The Fathers' View, as I'll call it, in a nutshell, is this:

> Dissensus happens only as brief episodes–epistemic accidents at it were–caused by error or blocked access to the full set of available evidence. They belong to the context of discovery and leave no enduring mark on science. Science would remain essentially unchanged and intelligible if dissensus disappeared from the history (actual processes) of science.

ii. Dissensus as the permanent state of science

In the Fathers' View, there is a reason why consensus is the necessary end-state of any scientific inquiry: the scientific method demonstrably tracks truth, and scientists are professionally committed to following the scientific method. In 'post-Legendary' philosophy of science, both assumptions are put in question, and through the lens of constructivist/historicist sociology of scientific knowledge, consensus in science, far from being the normal outcome of inquiry, appears as socially imposed discipline on a state of permanent cacophony – it is nothing but 'procedurally enforced consensus'[5] which bears no relation to the rational convergence to truth which the Fathers envisioned. The Bad Sons' View can be summarized thus:

> Dissensus is the natural state of science in the making, and yields only to political force exerted by one camp, deploying a mix of rhetorical and institutional maneuvers, either at the time of discovery (science in the making) or at the time of evangelical reconstructions (at the stage of pedagogy, for both lay and professional audiences).

iii. Dissensus as essential but transitory stages of the growth of scientific knowledge

As is well-known, the pendulum has swung back and the main trends in contemporary philosophy of science seek to integrate the justified objections to the 'Legendary' picture of science and in particular to the Fathers' View of dissensus as accidental, while holding onto a rationalistic conception of science. This had led to a rehabilitation of consensus, conjoined with a novel respect for dissensus. The resulting Good Sons' View goes something like this:

> Dissensus is a fact of life. It reflects the imbalance between the complexity of the world and our cognitive abilities: science is hard and makes consensus a protracted process. It is therefore an enduring feature of developing science,

3 The philosophers-scientists who gave birth to modern science.

4 Peter Machamer, Marcello Pera, and Aristides Baltas, *Scientific Controversies*. New York, Oxford: Oxford University Press 2000, p. 50.

5 Stephen Fuller, "The Elusiveness of Consensus in Science", in *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1986, vol. 2, p. 111.

which leaves a trace in its crystallized forms. It also has instrumental value (as an antidote to dogmatism and error, as a heuristic device …). Still, what scientists aim for, and eventually achieve, is consensus, either of the Fathers' type or pluralistic.

This view seems to point to two possible resolutions of the problem, which I want to acknowledge while denying that they actually settle the matter:

(i) *The instrumental value of dissensus*. The dialectical tradition stemming from Aristotle gives critical dialogue a crucial (in fact in some versions constitutive) role in rational inquiry. Criticism is the key methodological maxim in Popper's critical rationalism. But dissensus is not criticism, as we saw: it is a state of enduring plurality of views. Still, dissensus can, and often does lead to sustained criticism; it is thus an important, perhaps indispensable error-correcting mechanism that keeps science on the right track. There is yet a second line of thought, which runs from Mill to Feyerabend,[6] which grants dissenting opinions a protective role against not only error but dogmatism, smugness and tyranny: even a correct theory gets to lose some of its intellectual and practical virtues if it remains unchallenged. I have no objection to these views, except if they are presented as final: there is much more to dissensus than being the midwife of consensus or the guardian of honest thinking.

(ii) *Pluralistic consensus*. It might seem that an extra step is taken when we accept the possibility that several theories applicable to the same realm of nature can be simultaneously correct without being inter-reducible. But this is getting the dialectic back to front: scientific pluralism is, in the main (there are conflicting versions sharing the label), a purported solution by elimination to the problem of dissensus. The thought is that different perspectives on a complex domain yield different theories, all of which can be true or faithful representations at the same time: by letting go of the Fathers' demand for a unique all-encompassing theory, we can re-interpret cases of apparent enduring dissensus as two-tier consensus: consensus on each of the co-existing theories, and consensus on the legitimacy of their co-existence. No trace of dissensus left. Now there may well exist among conceptions of scientific pluralism some that *also* countenance genuine dissensus – such a theory would in fact have my own preference. But the mere move to pluralism is not enough to settle the issue of dissensus; in fact, it tends to push it under the rug.

6 Elizabeth Lloyd, "Feyerabend, Mill, and Pluralism", in: *Philosophy of Science* 64, Suppl., 1997, pp. 396-407.

2. The (merely) social turn

Up to this point, I have tacitly assumed an individualist perspective, that of the Fathers. It should be mentioned in this respect that individualism does not force the acceptance of the Consensus Thesis. Nicholas Rescher, for one, seems to have anticipated some of Solomon's conclusions: he defends the idea that there is nothing intrinsically rational about seeking consensus, and moreover that dissensus is the inevitable consequence of the fact that "equally rational inquirers placed in different experiential situations will come up with variant answers to the question of how things are. […] Reason is exercised from particular and differentiated places"[7].

Rescher shares with the pro-consensus classical thinkers a purely individualist conception of the epistemic situation in science, which may be summarized thus:

> [Ind] The Scientist inspects Nature and establishes that Nature has feature F.

In the last twenty years however, philosophy of science has taken a social turn and now tends to favor the following picture:

> [Com] The Community of scientists inspects Nature and establishes that Nature has feature F.

which can be construed in several rather different ways, actively explored in social epistemology. First, the Fathers' understanding, which is also Popper's and Rescher's and antedates the social turn, places the collective moment in the beginning, during the discovery phase: tasks may be parceled out for greater efficiency, and/or critical exchanges between participants can contribute to the emergence of truth. From then on, a 'Principle of Individual Recapitulation' brings the result of this collective labor under individual jurisdiction: some member of the collective at least can, in principle, gather the entire evidence and master all the inferential steps leading to the desired conclusion, in effect severing all dependency links with her collaborators. As far as I can see, this reading of [Com] presents with respect to [Ind] no difference that would be relevant to the issue of consensus.

By contrast, the social turn is really taken when the Principle of Individual Recapitulation is cast away: social in its genesis, scientific knowledge remains social at all stages. Our principle is then open to two main interpretations, according to our understanding of the nature of the relevant group:

> [C-Com] The Concrete Community of scientists inspects Nature and establishes that Nature has feature F.

---

7 I confess to not having read Rescher's book: *Pluralism: Against the Demand for Consensus*, Oxford: The Clarendon Press 1993. This is a gloss by a reviewer: David Archard, "Review: The Morality of Pluralism, by J Kekes, and Pluralism, by N. Rescher", in: *The Philosophical Quarterly* 45, 1995, pp. 400-403.

> [A-Com] The Abstract Community of scientists inspects Nature and establishes that Nature has feature F.

How best to characterize and contrast concrete and abstract communities is a topic of lively discussion in social epistemology. The first roughly correspond to a collective of active individuals with interpersonal connections, possibly somewhat remote or indirect, but with a sense of constituting a self with specific abilities, competencies and responsibilities.[8] The second sort lacks these connections and this sense of self. Of course there is a continuum ranging from, say, a pair of scientists co-signing a paper to the (fuzzy) set of all scientists since Galileo, and where to usefully draw any sort of line depends on what issue one is trying to address. Regarding consensus and dissensus, we want to understand what it means for there to be a consensus or not, given that, or insofar as we have ceased to believe that individuals are the sole loci of scientific beliefs. In the Fathers' paradise, it was straightforward to attribute consensus in a community, anywhere along the concrete/abstract line: there was a consensus on F whenever every member, or almost every member of the community believed (or was justified in believing) F. One way to go in the present setting might be as follows: (1) Define a concrete community as one to which there are good reasons to attribute beliefs of its own. It might be, for example, a group with such a rich background of shared beliefs and practices that, given a feature F which the group believes to belong to Nature, any of its members can convincingly fight back any challenge to the effect that she doesn't justifiably believe that Nature has feature F, even as she might rely on the rest of the group to fill out the meaning of F or the reasons to think that Nature has F. (2) Now take any abstract community C in science with a concern for F, and say that there is a consensus regarding F if every concrete community within C has a collective justified belief that Nature has F.

There are two connected problems with this proposal. The first is that of dealing with the dead. Do we include them or not? It's hard to include them in concrete communities, as long at least as we think of them roughly along the lines suggested above: dead people are not usually regarded as active, communicating, committed, responsible entities. On the other hand, by leaving them out we deprive concrete communities of important resources: it's not only during the discovery phase that most if not all scientists rest on the shoulders of the giants who preceded them. This dilemma leads us directly to another problem, which is actually the same problem in a wider perspective. It is the problem of acquired scientific knowledge. When Sara comes home and investigates the beer situation, she is no doubt relying on background knowledge regarding the nature of beer, the geometry of bottles, the medium-term stability of refrigerators, the habits of the household and so

8 Among the most discussed proposals, Margaret Gilbert's 'plural subjects', of which established groups are an instance, fill the bill of my concrete communities; see her *Sociality and Responsibility*, Lanham: Rowman & Littlefield 2000. I remain uncommitted here to this or any other specific proposal.

forth. This knowledge is committed to memory, and in this and similar cases it is in fact integrated in a set of mostly automatic behavioral patterns in such a way that at no time does Sara (the person, not her neurons) have the need to refer to it; in other cases Sara will also consciously refer to stored factual knowledge. But it is still always Sara, carrying her own resources inside her, who is facing the world and asking whether there is any beer left at home.

On an unreflective reading of [Com], when a scientist inquires whether Nature has feature F, like Sara he has assimilated all the required background knowledge so that it is now 'ingrained' , 'incorporated' in his memory and his thought patterns: he approaches Nature the way Holmes approaches the case of the speckled band, with his bare cognitive and sensory organs. Now I believe this is already utterly off track, but I will let it be for a moment and go to the other side of the social turn.

Let us first consider the case of a concrete community, and ask how it is supposed to 'inspect' Nature. Does it come to it with its bare (cognitive and sensory) organs? What are they? Where are they? These are the questions which the area of 'distributed cognition' proposes to answer. It may in fact succeed: after all, once we have secured some workable notions of collective belief and collective intention, the problem of making sense of collective acquired scientific knowledge may be soluble. So, although we are not in better shape than in the individual case, at least we might not be in distinctly worse shape.

Things do get worse when we move to abstract communities. While in a concrete community, there are naturalistic, causal connections between members, such that somewhat like bees which build a hive and collect honey following elaborate collective strategies, scientists build a common home made of shared practices and tools, and collect communal knowledge, these natural links and ongoing processes are lacking in an abstract community. And presumably this is what makes it difficult to grant abstract communities the capacity of forming beliefs and intentions, and therefore of having and deploying established knowledge, except of a summative (and thus impotent) kind. Thus, I submit, although we may perhaps accept [Ind] and (just barely) [C-Com] by assuming that the 'inspecting' and 'establishing' occur in the presence of stored resources, internal, as the case may be, to the individual or to the concrete community, we cannot go further: [A-Com] is not intelligible as it stands (except of course in the metaphoric sense with which we are all familiar).

3. Public scientific knowledge

As Popper long ago,[9] and more recently Alexander Bird[10], have proposed, the role which is attributed to the concept of (scientific) knowledge in various contexts cannot be filled by a single concept: besides the concept of subjective knowledge which is deployed in epistemology, of both the traditional, individualist and the contemporary, social kinds, there is a need for something which Popper proposes to call 'objective knowledge', and Bird 'social knowing'. There are differences between the two notions, and Popper's has been shown to run into serious difficulties.[11] They respond nonetheless to the same urge to sever, or to at least seriously weaken the link between people's states of minds and *truly* social epistemic states. I propose to call *public scientific knowledge* the set of publicly available counterparts of individually and socially held representations, methods, instrumental and intellectual skills which together constitute the competence of scientists or (concrete) scientific communities. It is not in the purview of this paper to defend this working definition or to propose a better one: I mean it to fill the role of the disambiguated sense of 'science' in which it becomes (again) quite legitimate to talk about what science knows and doesn't know at a certain moment in time, what it's confident about and what it's unsure of, and so on.

The basic epistemic situation for public scientific knowledge is this:

> [Pub] As a result of a concerted effort of individuals and teams, public scientific knowledge, initially in state S1, moves to a new state S2 which includes an attribution to Nature of feature F.

But it now appears that public scientific knowledge (henceforth Science for short) has a life of its own, and that individual scientists as well as communities of scientists are not just facing Nature, but Nature *and* Science. At this point, my earlier characterization of the individual situation becomes untenable. In its stead we might consider something like this:

> [Ind*] The Scientist inspects Nature and Science [at stage S1] and establishes that Nature has feature F, as expressed and understood (by her) in the context of Science [at stage S2].

It will be objected that Nature and Science are not on the same footing: Science is about Nature, while Nature is about nothing at all; Science acts only via the scientist's cognitive apparatus, while Nature has original causation. The scientist inspects Nature in part by inspecting Science, but not Science by inspecting

9 In: Karl R. Popper, "Epistemology Without a Knowing Subject", ch. 3 of his *Objective Knowledge*, Oxford: Oxford University Press, 1972, pp. 106-152; first publication 1968.

10 Alexander Bird, "Social Knowing. The Social Sense of 'Scientific Knowledge'", in: *Philosophical Perspectives*, 24, 2010, pp. 23-56.

11 L. Jonathan Cohen, "Some Comments on Third World Epistemology", in: *The British Journal for the Philosophy of Science* 31, 1980, pp. 175-180.

Nature (at least, not in a straightforward way). Finally, but this is indicated, Science changes in the process while Nature remains (essentially) unchanged. On the other hand, if we are willing to consider for a moment what scientists, from graduate students to Nobel laureates, actually do from dawn to dusk, we must I think concede that 'inspecting Nature and Science' is a less inadequate characterization than, say, 'inspecting Nature in the light of Science'. While sounding more respectable, this way of putting it obliterates the fact that far from only helping the scientist figure out what's there or how things work the way a torch helps Holmes detect a speck of dust on a pew, science is so to speak a complex and ever changing cathedral which she is constantly learning and relearning to navigate. The larger point is that Science (public science) is not delivered at the cognitive door of the scientist for instant plug-in (or to use another metaphor, one familiar in the discussion of education, it does not come, ready made, on a conveyor belt reaching straight into the scientist's mind), it is deciphered and interpreted by the scientist.

A more serious objection to this characterization is that, according to the social conception of science, the individual scientist is helpless if left entirely to her own device. So [Ind*], though improving on [Ind], must go. We are lead to one of the following:

> [C-Com*] The Concrete Community of scientists inspects Nature and Science [at stage S1] and establishes that Nature has feature F, as expressed and understood (by the Community) in the context of Science [at stage S2].
>
> [A-Com*] The Abstract Community of scientists inspects Nature and Science [at stage S1] and establishes that Nature has feature F, as expressed and understood (by the Community) in the context of Science [at stage S2].

But [C-Com*] is reasonable only as a schematic description of a subjective, perishable, parochial process: the social-epistemic dynamics of a given, historically situated group of people. Although Science emerges from such dynamical processes, it doesn't reduce to any one of them. Replacing the concrete communities by an abstract one, which is the move implicitly commended by the traditional perspective, leads us to [A-Com*], which makes even less sense than [A-Com]: [Pub] must take its place.

With this dual description of the epistemic process, social/subjective, [C-Com*], and public, [Pub], we are at last in a position to untie the dissensus knot. Recall the starting point: it did not seem possible to reconcile the fact that science aims for truth and the fact that science need not aim for consensus without twiddling with the concept of truth, which I have disallowed.

Aiming for truth I have first proposed to construe in terms of what I have called the basic epistemic situation. Going social has meant locating the agency in concrete communities: those are in the business of finding out whether Nature has feature F. But science cannot be described solely in terms of concrete com-

munities: there is a public sense of determining that Nature has F, which involves Science (at successive stages).

The drive towards consensus, the urge to eliminate the 'irritation' (Peirce's term) caused by diverging conclusions regarding F, occur within concrete communities: this is where the classic picture applies. But F seldom, if ever, concerns just one concrete community.[12] When two such communities inspect Nature and Science to find out about F, while we may (for simplicity's sake) assume that they would get from *Nature* the same answers if they asked the same questions, what they draw from *Science* (at stage 1) may be non-identical interpretations S_1' and S_1", leading to different sets of questions and different ways of integrating Nature's answers in S_1' and S_1", eventually leading to S_2' and S_2" which include opposite conclusions regarding F. The two communities' labors are projected onto (public) Science, which thus exhibits the public counterpart of a dissensus regarding F. And although it is possible of course that at some later point, some concrete community will take up as its task to relieve the tension, that tension is not felt by Science, which doesn't feel anything. Time may well pass before a resolution is proposed, either through a deliberate effort to settle the matter, or as a side effect of some development in another area. And if a sufficiently long time elapses without a resolution, the chances are that F will have fallen out of the conceptual vocabulary of science anyway.

In a nutshell, the standard view, re-affirmed by Richardson, is correct when restricted to concrete communities and weakened to leave space for more pressing rational constraints. But as Solomon is right to stress, dissensus can appear and endure on the public plane, when different communities, while interested in the same feature of Nature, follow divergent social-subjective trajectories without feeling any compulsion to blend (or having any way of blending) into one concrete community in charge of resolving their difference (of which they may not even be aware). So that public science contains theories that are plausibly understood as contradictory, and no mechanism to uproot them, neither as quickly as possible after they appear, nor in the fullness of time.[13] Or again: there exist two kinds of dissensus, social-subjective and public; the first is destined to disappear (in principle if not in fact), the second is not.

12 This is where we need to be clearer on what a concrete community is. It need not be a school of thought, in fact, it better comprise competing schools, for all the Mill-Popper-Feyerabend reasons. On the other hand, it need not be the entire population of specialists of a given area, which may well be divided in communities which essentially don't talk to each other, who may not even be clearly aware of the others' existence (for example because they belong to different superordinate disciplines).

13 This was one of Cohen's main worries regarding Popper's objective knowledge (*ibid.*). I think my proposal puts it to rest.

4. The picture picture of science abandoned and the normative status of dissensus established

What we have at best secured at this juncture is an argument against the vanishing status of dissensus: we can see why dissensus occurs and why we have no strong reason to believe it should eventually disappear, but we are in want of an argument showing why enduring dissensus should not in fact be exceptional or at least rare, let alone in what sense it could be a norm, as my title suggests.

It would be interesting to be able to actually count cases of dissensus and compare the figure to the number of cases of consensus, but the prospects of a counting or measurement method are slim. We must content ourselves with the converging impressions of a number of authors with extensive historical knowledge, such as Solomon, Rescher, Laudan, Kitcher, and the many scholars of scientific controversies, who emphasize the non-exceptional character of dissent.

In the social-subjective sense of dissensus, we can first agree with Solomon on a negative construal of the norm of dissensus: it may be better, all else being equal, from a rational standpoint, to remain at odds with an established theory and thus, so to speak, follow a dissensus rather than a consensus strategy.[14] This complements the positive sense in which dissensus is a norm: it favors originality and the debugging of errors in established theories; by introducing diversity, it increases the probability of solving problems and overcoming impasses.[15]

But dissensus as a feature of public scientific knowledge is also, I submit, a norm, in the sense where it occurs habitually and not as a fluke, and in the sense where it contributes to the scientific enterprise.

Why would dissensus occur habitually? There is no space left for a detailed argument, so I'll proceed sketchily. I see two structural features of science which generate dissensus. The first is what I'll call for brevity's sake the *fish-scale effect.* The phrase is due to Donald Campbell,[16] who likens the fit between science and nature to that of the coat of scales on the fish's body: each scale fits tightly the patch of skin it protects, the scales overlap thus providing full coverage of the animal, yet together the scales do not constitute a continuous, tight-fitting cover. Now if something like the fish-scale effect operates in science, then dissensus seems bound to arise, as different people will choose different distributions of contact points. One way of fleshing this out might be to think of a contact point as the founding problem or key phenomenon of some research program. In the

14 As Solomon writes: "it is not important for a scientist to get the opposition to convert or die. […] What matters is that a scientist develop empirical successes – especially unique empirical successes – in their own theory." "Responses to critics", in: *Perspectives on Science* 16, 3, 2008, p. 282.

15 Scott E. Page, *The Difference*. Princeton: Princeton University Press 2007, chapter 6.

16 Donald T. Campbell, "Ethnocentrism of Disciplines and the Fish-scale Model of Omniscience", in: Muzafer Sherif and Carolyn Wood Sherif (Eds.), *Interdisciplinary Relationships in the Social Sciences*. Chicago: Aldine 1969, pp. 328-348.

favorable case, a theory is developed which provides a satisfactory solution to that problem or a scientifically irreproachable account of that phenomenon, but the theory's fit to other phenomena in the vicinity of the starting point turns out to be less satisfactory. Another team might take one of those as its starting point and develop another theory which is bound to conflict with the first.

An escape from this predicament might exist if we could establish that the contact points are objectively determined, that they are, so to speak, marked on the body of Nature (some dual image, of sorts, of the notorious 'joints' of Nature that mark out the true universals) – but I am not enough of a metaphysician to see how that could be done.[17] But at any rate, we can put the proposal thus:

> [Fish-scale] If science provides only local maps of Nature, then dissensus is inherent in its development.

The second structural constraint which I suggest forces dissensus as a normal aspect of science is what I call radical incompleteness. Contrary to what the fish-scale model might suggest,[18] the coat of scales never comes close to covering the whole beast, even in the infinite limit of the end of history. Science does not gradually 'fill out the picture', it doesn't 'complete the puzzle'. There isn't a finite (albeit gigantic) set of empirical facts out there that are gradually brought to light and integrated into ever more encompassing theories. There are indefinitely many ways of chunking Nature, and indefinitely many sets of questions to ask, and although science makes steady progress, it never gets any closer to exhausting its general agenda.

If this is indeed the case, then it seems that the gradual stabilization which occurs in fields of limited scope does not extend to science as a whole: fields keep growing, boundaries shift, empirical facts continue to arise, science undergoes tectonic reconfigurations. In the process, dissensus, smoothed out of one end of the rug, reappears at another: newly formed theories arising in the novel context conflict with established ones, and between themselves.

Let me state this second structural constraint in conditional form:

> [Incompleteness] If science is radically incomplete, then dissensus is inherent in its development.

Let me try to piece things together, tying well-known phenomena which I have hardly touched upon and those which I have tried to pin down. In both social-subjective and public spaces, some permanent dispositions tend to create dissensus, just as other dispositions tend to create consensus.

In social-subjective space, consensus-increasing factors are: the demand for consistency and the need for expediency (which encourages communities to adopt a 'satisficing' view of consensus); the dissensus-increasing factors are: the quest

17 Of course it remains open to someone, say a radical scientific realist, to reject the metaphor altogether, in which case the issue doesn't arise.

18 I doubt it was part of Campbell's thought.

for originality, the search for resolution of anomalies and the underlying mistaken assumptions, the division of labor leading to divergent mindsets and commitments, the inevitable diversity in starting points (no two scientists, and no two communities having the exact same initial set of data, assumptions and tools[19]) and finally perhaps sheer intellectual curiosity. As the Fathers and the Good Sons insist, concrete communities do have consensus as a goal, or perhaps as a hope: they do care and are not comfortable with enduring dissent. But they cannot have as their sole goal the suppression of this source of discomfort.

In public space, over and above the traces left by the communities, which are so to speak the projection of the factors at work in the social-subjective space, there are structural factors which operate independently. I have not examined the case of those which tend to increase consensus. Perhaps there is a general argument to the effect that public science tends toward greater consensus for structural (*e.g.* transcendental) reasons, or that a unifying theory will eventually be seen to provide a general framework within which pockets of dissensus will be reduced one by one; but I have no such argument to offer. On the other hand, there is a sociological factor at work, which is the operation of what may be called the scientific police: its job is to minimize the influence of heterodox voices at all levels of public scientific life. For dissensus-increasing factors, in reverse order, the sociological/institutional one I can think of is the pluralistic organization of academia: each institution attempts to create its own niche. The two epistemic factors I have suggested are the fish-scale effect and the incompleteness effect.

The resulting picture is a far cry from either the Fathers' or the Sons' views. Instead of having a space of public science tending towards monophony, a great book of truth which all scientists eventually subscribe to, or else a space of forever conflicting theories, what we see, or so I suggest, is a dynamic field where areas of varying levels of consensus develop, contract and expand under the combined effect of consensus and dissensus-favoring factors. The normative monopoly of consensus has been displaced in favor of a more symmetric distribution, at both the social-subjective level, as Solomon argues, or at the public level, as I have tried to show.

Finally, I have claimed that the normative status of dissensus is also instrumental. Again, at the social-subjective level the point has been abundantly made by Mill, Popper and other authors, some now active in social epistemology. But dissensus also favors the public or objective life of science, endowing it with resilience in the face of an uncooperative Nature. The reason is that when a theory

19 According to Andrew Lugg, (i) scientists have different access to data, so that (ii) they are bound to come to different conclusions and (iii) there is no way of eliminating 'access differences' and hence no way of eliminating disagreement in science without adversely affecting one or another aspect of the scientific enterprise itself. "Thus, it would be a mistake to think that disagreement among scientists is incidental to science". Andrew Lugg, "Disagreement in science", in: *Journal for General Philosophy of Science* 9, 2, 1978, pp. 276-292.

collapses, the existence of an alternative in public space, ready to take over and, so to speak, to hit the ground running, allows science to quickly overcome its failure and continue the inquiry, rather than abandon the project until sometime, somewhere, a new line of thought emerges, if it ever does. Note how important it is not to link the alternative theory to a particular concrete community (for example, as in Kuhn, a rival, younger school, or more generally, one engaged in a controversy with the one which has just collapsed): as long as it is poised for uptake by a scientist (it may be the work of a dead and forgotten author, as in one of Bird's examples), it remains a live possibility. This idea is by no means original: it is a standard theme in evolutionary epistemology à la Hull[20]– keeping alive a diversity of genotypes protects a population from going extinct and helps it move to a more hospitable niche. I said that Nature was the uncooperative one, but the evolutionary scheme suggests a more complex picture: theories don't only face falsifying data, they also co-exist with other theories, and can be pushed out of the picture when irreconcilable differences arise. As this example shows, and that will be my concluding remark, there are many important issues, the status of dissensus being one, where taking the social turn only goes half way: as Popper and Bird insist, it must be completed with the public turn.

UFR de philosophie et sociologie
Université Paris-Sorbonne
1, rue Victor Cousin
75230, Paris cedex 05
France
daniel.andler@paris-sorbonne.fr

20 David L. Hull, *Science and Selection*. Cambridge: Cambridge University Press 2000.

Index of Names

Not included are: Tables, References and Figures

H. Andersen et al. (eds.), *New Challenges to Philosophy of Science*,
The Philosophy of Science in a European Perspective 4,
DOI 10.1007/978-94-007-5845-2, © Springer Science+Business Media Dordrecht 2013