European Studies in Philosophy of Science

Uskali Mäki
Ioannis Votsis
Stéphanie Ruphy
Gerhard Schurz *Editors*

Recent Developments in the Philosophy of Science: EPSA13 Helsinki

Springer

European Studies in Philosophy of Science

Volume 1

This new series results from the synergy of EPSA - European Philosophy of Science Association - and PSE - Philosophy of Science in a European Perspective: ESF Networking Programme (2008–2013). It continues the aims of the Springer series "The Philosophy of Science in a European Perspective" and is meant to give a new impetus to European research in the philosophy of science. The main purpose of the series is to provide a publication platform to young researchers working in Europe, who will thus be encouraged to publish in English and make their work internationally known and available. In addition, the series will host the EPSA conference proceedings, selected papers coming from workshops, edited volumes on specific issues in the philosophy of science, monographs and outstanding Ph.D. dissertations. There will be a special emphasis on philosophy of science originating from Europe. In all cases there will be a commitment to high standards of quality. The Editors will be assisted by an Editorial Board of renowned scholars, who will advise on the selection of manuscripts to be considered for publication.

More information about this series at http://www.springer.com/series/13909

Uskali Mäki • Ioannis Votsis • Stéphanie Ruphy
Gerhard Schurz
Editors

Recent Developments in the Philosophy of Science: EPSA13 Helsinki

Springer

Editors
Uskali Mäki
University of Helsinki
Helsinki, Finland

Stéphanie Ruphy
Pierre Mendés-France University
Grenoble, France

Gerhard Schurz
DCLPS
Heinrich-Heine University Duesseldorf
Duesseldorf, Germany

Ioannis Votsis
DCLPS
Heinrich-Heine University Duesseldorf
Duesseldorf, Germany

Philosophy Faculty
New College of the Humanities
London, UK

ISSN 2365-4228 ISSN 2365-4236 (electronic)
European Studies in Philosophy of Science
ISBN 978-3-319-23014-6 ISBN 978-3-319-23015-3 (eBook)
DOI 10.1007/978-3-319-23015-3

Library of Congress Control Number: 2015950931

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

The European Philosophy of Science Association [EPSA] was established in 2007 to promote the philosophy of science on the European continent and beyond. One major way this is achieved is through the Association's biennial conference, which brings together hundreds of philosophers of science from numerous countries, working on a variety of sub-fields. The 2013 conference took place at the University of Helsinki, and was organized by the Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences [TINT]. For the programme and other details, see www.helsinki.fi/epsa13/.

As an indication of the growing interest in EPSA and its activities, the EPSA13 Programme Committee, chaired by Stéphanie Ruphy and Gerhard Schurz, received submissions for 391 contributed papers and 21 symposia. About one third of the proposed contributed papers and one half of the proposed symposia were accepted. This resulted in a programme with 206 speakers in 47 sessions, of which 12 were symposia and 3 were invited keynote plenaries.

After each EPSA conference a corresponding Proceedings is put together. Its aim is to showcase written up versions of some of the very best work presented at that edition of the conference. The present volume contains twenty-nine peer-reviewed contributions sourced from the 2013 meeting of EPSA. The purpose of this preface is not to provide a detailed overview of each contribution – there are simply too many to do them justice here – but rather to give the reader a little foretaste of the kinds of topics on offer.

The twenty-nine contributions can be classified under the following coarsely-grained topic headings (their frequency stated within brackets): general philosophy of science (x7), philosophy of physics (x5), social epistemology (x4), philosophy of probability (x3), philosophy of chemistry ($x2$), philosophy of medicine ($x2$), philosophy of the social sciences and humanities ($x2$), philosophy of biology (x1), philosophy of mathematics (x1), philosophy of psychology (x1) and philosophy of science and public policy (x1). Of course this classification is imperfect, as some papers deal with more than one of the foregoing headings. Moreover some headings have not even made our list. For example, some papers have a discernible historical vein, though not discernible enough to warrant the label 'history of science papers'. In

spite of the limitations of our classification, we feel that it is adequate for the task at hand, namely to inform the reader about the distribution of topics in this collection.

A more finely-grained identification of topics demonstrates the rather broad distribution of interests: gender-specific medicine (Amoretti and Vassalo), laws of nature, partial structures and paraconsistent reasoning (Andreas), science funding (Avin), quantum field theory (Bain), fictions, explanation and thermodynamics (Bangu), causality and physics (Bartels and Wohlfarth), values and underdetermination (Bueter), debate dynamics and truthlikeness (Cevolani), judgment aggregation and wisdom of crowds (Feldbacher), Carnap's inductive logic and explications (French), local causality and Bell inequality (Gábor), concepts of emotional expression (Glazer), Bertrand's paradox and interpretations of probability (Gyenis and Rédei), chemical kinds (Hendry), manipulation, intervention and explanation (Kästner), objectivity and values in indigenous studies (Koskinen), causality, complexity and abstraction (Kronfeldner), clinical trials (Nardini), logical empiricism and structural realism (Neuber), causality and decoherence (Samaniego), measures of coherence and truthlikeness (Schippers), robustness analysis and evidential diversity (Schupbach), explanation, idealisation and reduction in quantum chemistry (Seck), scientific representation, fiction and denotation (Suárez), biological essentialism and species pluralism (Talpsepp), realism, scientific progress and verisimilitude (Tambolo), meta-induction and wisdom of crowds (Thorn), rational choice theory and normativity (Tiefensee) and definitions of chaos (Zuchowski).

We would like to extend our warmest gratitude to a number of individuals and organisations. First and foremost we would like to thank the authors whose contributions are, needless to say, the main attraction. We would also like to thank the referees most of whom served on the EPSA13 programme committee. In alphabetical order: Alban Bouvier, Ellen Clarke, Michael Cuffaro, Foad Dizadji-Bahmani, Isabelle Drouet, Kevin Elliott, Mathias Frisch, Sven-Ove Hansson, Stephan Hartmann, Janet Kourany, Bernd Lahno, Endla Lõhkivi, Kerry McKenzie, F.A. Muller, Nancy Nersessian, Wendy Parker, Tom Pashby, Helmut Pulte, Alexander Reutlinger, Bryan W. Roberts, Samuel Schindler, Sonja Smets, Katie Steele, Michael Stöltzner, Mauricio Suárez, David Teira, Charlotte Werndl and Jan Wolenski. Additionally, we would like to thank the local organising committee of EPSA13 for all the effort put into an excellently organised conference, as well as the EPSA steering committee for facilitating our work and for putting the ball in motion. Finally, we would like to thank the good people at Springer who helped make this volume a reality. Three names, in particular, deserve to be mentioned: Christi Lue, Madhuriba Subarayalou and Ties Nijssen. We very much hope that the readers will find the contents of the Proceedings edifying and inspiring, thereby making the efforts of the above mentioned all the more worthwhile.

Helsinki, Finland — Uskali Mäki
London, UK — Ioannis Votsis
Grenoble, France — Stéphanie Ruphy
Duesseldorf, Germany — Gerhard Schurz
January 2015

Contents

Part I
Truth and Semantics

Coherence and (Likeness to) Truth

Michael Schippers

1 Introduction

Methodological questions regarding the assessment of scientific theories are a central subject matter in philosophy of science. In this paper the focus is on a coherentist account to theory comparison. The core idea behind this account has recently been characterized by Peter Brössel (2013) as follows:

> [A] theory is a good theory given some observational data if and only if that theory coheres with the observational data, and [...] a theory is better than another theory given some observational data if and only if the first theory coheres more with the observational data than the second. (p. 594)

In what follows I want to consider the adequacy of this coherentist account with a special focus on probabilistic measures of coherence. More precisely, I want to answer the question whether probabilistic measures of coherence are proper assessment functions for comparing scientific theories (cf. Brössel 2013; Huber 2008). To this end, I will introduce some adequacy constraints each assessment function ought to satisfy in order to count as an adequate measure of the epistemic value of a scientific theory. Many of these constraints originate from discussions on the verisimilitudinarian account of the nature of scientific progress as advocated by Ilkka Niiniluoto (1984, 1987, 1999), Theo Kuipers (1987, 2000) and others. In a nutshell, supporters of this position claim that scientific progress is to be thought of as increasing verisimilitude or approximation to the truth.[1] In this sense, "such

[1] In this paper I will use the terms "verisimilitude" and "truthlikeness" as synonyms.

M. Schippers (✉)
Department of Philosophy, University of Oldenburg, 26111 Oldenburg, Germany
e-mail: mi.schippers@uni-oldenburg.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_1

theory-changes as the transition from Newton's to Einstein's theory are progressive because, although the new theory is, strictly speaking, presumably false, it is estimated to be closer to the truth than the superseded one" (Cevolani and Tambolo 2013, 922).

In keeping with the verisimilitudinarian account, Sect. 2 discusses two important factors for scientific theory comparison. In Popper's words, "science does not aim, primarily, at high probabilities. It aims at high informative content, well backed by experience" (1968, p. 399). Therefore, I am going to show that existing coherence measures can account for both these factors.

Section 3 then elaborates on a list of further adequacy constraints for proper assessment functions. The focus of this discussion will be on constraints that compare two scientific theories with respect to some observational data. Thus, the question to be answered by means of coherentist assessment functions is not whether t is a good scientific theory in the light of observational data, but whether t is better than another theory t' in the light of these data. The upshot will be that although there are coherence measures that satisfy the vast majority of constraints, all of these measures suffer from a serious drawback: according to all these coherentist accounts, a false theory can never be more epistemically valuable than another (true or false) theory.[2] To eliminate these shortcomings, Sect. 4 introduces a recipe for *two-sided* coherence measures that allow for reasonable assessments of falsified theories. The basic idea underlying these measures is to think of coherence as a *balance* between the verified parts of a theory and the evidence on the one hand, and the falsified parts of the theory and the *inverse* of the evidence on the other.[3]

The last section is then devoted to a vindication of a coherentist approach to scientific theory comparison that shows that coherence is a means for achieving (estimated) truthlikeness. More precisely, focusing on *conjunctive* theories (cf. Cevolani et al. 2011) the following implication will be shown to hold for our favored two-sided coherence measure: if a theory t better coheres with the available observational data than another rival theory t', then t is also assigned a higher degree of estimated truthlikeness than t' according to a large number of existing accounts to truthlikeness.

Formal preliminaries: Let $\mathcal{L}_n$ be a finite propositional language with n atomic propositions $p_1, \ldots, p_n$ and the standard connectives $\neg, \wedge, \vee$. Arbitrary formulae of $\mathcal{L}_n$ will be denoted by lower-case Greek letters φ, ψ; sets of such formulae will be denoted by upper-case Greek letters Γ, Σ. As usual, $\top$ and $\bot$ denote a tautology

[2]The first philosopher to propose a definition of verisimilitude that suffered from a similar drawback is Karl Popper. Miller (1974) and Tichý (1974) independently proved that Popper's original definition Popper (1963, 1972) suffered from a logical flaw so that on his account a false theory can never be closer to the truth than another true or false theory. For a comprehensive survey of subsequent developments see Oddie 2008, Niiniluoto 2011 and Zwart 2001.

[3]See Sect. 4 for details.

and a contradiction, respectively. A *literal* $\pm p_i$ is either an atomic formula or its negation. A *probability function* over $\mathcal{L}_n$ is a real-valued function such that for all $\varphi, \psi \in \mathcal{L}_n$:

(i) $\Pr(\varphi) \geq 0$
(ii) $\Pr(\top) = 1$
(iii) $\Pr(\varphi \vee \psi) = \Pr(\varphi) + \Pr(\psi)$ if $\models \neg(\varphi \wedge \psi)$.

Finally, a probability function is called *regular* if only contradictions are assigned the minimum probability 0. In what follows, Pr is always assumed to be regular.

2 Coherence and Information

In the context of theory assessment, formal characterizations of coherence are utilized in order to render precise the degree of coherence between a scientific theory t on the one hand and the available evidence e on the other.[4] An informal characterization of this notion of coherence is given by Bonjour (1985), Harman (1986), Lehrer (1974) and others. The main idea is that coherent propositions "hang together" or "dovetail" with each other. In recent years probabilistic (aka Bayesian) measures of coherence have become increasingly popular in philosophical circles. The following families of coherence measures can be distinguished:

Deviation from probabilistic independence: This family of coherence measures draws upon the idea of coherence as a deviation from probabilistic independence. In order to quantify the degree of deviation, the following coherence measure has been proposed:

(1) $\mathcal{D}(t, e) = \Pr(t \wedge e) / \left[\Pr(t) \cdot \Pr(e)\right]$ (Schupbach 2011; Shogenji 1999)

Relative set-theoretic overlap: The idea underlying this family of measures is that a set of propositions is the more coherent the higher the probability that all of the sets' propositions are true given that at least one is. To spell out this intuitive idea in probabilistic terms, the following measure has been put forward:

(2) $\mathcal{O}(t, e) = \Pr(t \wedge e) / \Pr(t \vee e)$ (Glass 2002; Meijs 2006; Olsson 2002)

Average mutual support: According to Bonjour (1985), "the coherence of a system of beliefs is increased by the presence of inferential connections between its component beliefs and increased in proportion to the number and strength of such connections" (p. 98). This idea has most prominently been cashed out in probabilistic terms by Fitelson (2003, 2004) and Douven and Meijs (2007). According to

[4] The first formal approach to coherence is Paul Thagard's neural network-based *constraint satisfaction* model of coherence (Thagard 1989; Thagard and Verbeurgt 1998). A detailed comparison of Thagard's approach with probabilistic approaches is beyond the scope of the present paper.

their proposal, the strength of inferential relations ought to be determined by means of probabilistic measures of confirmation that have been discussed for some time within Bayesian accounts to philosophy of science (Carnap 1962; Earman 1992; Howson and Urbach 2006). Put formally, a probabilistic measure of confirmation is a real-valued (partial) function c assigning each pair (t, e), where t is a given theory and e denotes the given evidence, a real number $c(t, e)$ representing the degree of confirmation that e provides for t. Constraining attention to such pairs allows for a straightforward representation of the idea of coherence as mutual support[5]:

(3) $\mathcal{S}_c(t, e) = 1/2 \cdot \big[\, c(t, e) + c(e, t) \,\big]$ (Douven and Meijs 2007; Fitelson 2003)

Thus, in order to calculate the degree of coherence between t and e we simply average the mutual confirmation between both. To this end, we can choose from a large class of ordinally non-equivalent measures.[6] However, only the coherence measures based on the following confirmation measures have been explicitly endorsed in the literature:

$d(t, e) = \Pr(t|e) - \Pr(t)$ (by Douven and Meijs 2007[7])
$k(t, e) = \big[\Pr(e|t) - \Pr(e|\neg t)\big] \big/ \big[\Pr(e|t) + \Pr(e|\neg t)\big]$ (by Fitelson 2003)
$f(t, e) = \Pr(t|e)$ (by Roche 2013[8])

The crucial difference between d and k on the one hand and f on the other is that the former are sensitive to probabilistic relevance while the latter is not. That is, d and k quantify the degree of *incremental* confirmation, where evidence e incrementally confirms theory t iff $\Pr(t|e) > \Pr(t)$, while f focuses on the amount of *absolute* confirmation, where e absolutely confirms t iff $\Pr(t|e) \geq r$ for some threshold $0.5 < r \leq 1$.

Now we turn to our evaluation of Popper's insightful remark on the aim of scientific inquiry, namely the quest for theories with "high informative content, well backed by experience". Two classical representations of the degree of information of a theory t, discussed in seminal work by Bar-Hillel and Carnap (1953), are the following:

- $inf_R(t) = \log_2 \big[1/\Pr(t)\big]$
- $inf_D(t) = 1 - \Pr(t)$

[5]The general case is slightly more intricate: Fitelson's (2004) account as well as Douven & Meijs' recipe are applicable to sets of propositions. As such they take into account the mutual degree of confirmation for each pair of non-empty, disjunct subsets of the given set under consideration.

[6]For a comprehensive overview see Crupi 2015 and Crupi et al. 2007.

[7]Douven and Meijs (2007) consider further confirmation measures to be fed into their recipe, namely the *ratio-measure* (Horwich 1982; Keynes 1921) and the *(log-) likelihood measure* (Good 1984; Zalabardo 2009) of confirmation. However, $\mathcal{S}_d$ is their favorite explication of coherence. Accordingly, since an evaluation of all possible coherence measures is beyond the scope of the present paper, we restrict our attention to $\mathcal{S}_d$ and neglect all others.

[8]This measure has independently been proposed by Schippers and Siebel (2012, Reassessing probabilistic measures of coherence, Unpublished manuscript).

Both inf_R and inf_D are strictly decreasing functions of the marginal probability $\Pr(t)$ and satisfy the following constraints:

(1) $inf(\top) \leq inf(t) \leq inf(\bot)$ for all contingent theories t.
(2) If $t \models t'$, then $inf(t') \leq inf(t)$ for all theories t, t'.

Given a regular probability function, according to (i) a tautology is assigned the minimum degree of information while a contradiction $\bot$ is assigned the maximum degree of information.[9] The information content also varies with logical strength (ii).

Hence, a necessary condition each proper assessment function ought to satisfy is the following constraint (‡):

(‡) α is an adequate function for assessing theories, taking as input a given theory t and the available evidence e, iff there is a function f such that $\alpha = f\big[\Pr(t|e), inf(t)\big]$ and for all given theories t, t' and given evidence e the following conditions are satisfied:

(i) If $\Pr(t|e) = \Pr(t'|e) > 0$, then $\Big[\, \alpha(t,e) \geq \alpha(t',e)$ iff $inf(t) \geq inf(t') \,\Big]$.

(ii) If $inf(t) = inf(t')$, then $\Big[\, \alpha(t,e) \geq \alpha(t',e)$ iff $\Pr(t|e) \geq \Pr(t'|e) \,\Big]$.

As the following theorem reveals, all considered coherence measures are proper assessment functions in this sense.[10]

Theorem 2.1. *The coherence measures $\mathcal{D}$, $\mathcal{O}$, $\mathcal{S}_d$, $\mathcal{S}_k$ and $\mathcal{S}_f$ satisfy* (‡).

Thus, we conclude that all considered coherence measures satisfy at least this minimal constraint for proper assessment functions. The next section will focus on a list of adequacy constraints in order to discriminate between measures with regard to their adequacy for the purpose of assessing scientific theories.

3 A Constraint-Based Evaluation of Coherence Measures

This section elaborates on a more detailed evaluation of coherence measures regarding their adequacy for assessing the value of scientific theories in the light of given evidence. As was mentioned before, the main intuition to be explicated

[9] This latter property, sometimes called the "Bar-Hillel-Carnap semantic paradox" (Floridi 2004, p. 198), might seem curious at first sight. However, (iii) follows naturally from the assumption that t's information content is related to the amount of state descriptions precluded by t. A tautology precludes no state description at all and is consequently assigned the lowest possible degree of informativity. A contradiction on the other hand precludes every possible state description and is accordingly maximally informative. In this sense, Bar-Hillel and Carnap (1953) state that "a false sentence which happens to say much is thereby highly informative in our sense. [. . .] A self-contradictory sentence asserts too much; it is too informative to be true" (p. 229).

[10] For a subset of measures, a similar theorem is due to Brössel (2013). The proof of Theorem 2.1 is straightforward.

is that good theories should factor in both truth and informativity. In other words, a good scientific theory should *predict many things* and many of its predictions should *turn out to be true*.

Suppose that theory t only entails nearly tautologous predictions regarding the outcome of an experiment, while its rival theory t' gives a detailed forecast. In this case, it is highly unlikely that t is false, "simply because it tells us nothing, or very little" (Popper 1968, p. 399). Correspondingly, it might reasonably be assumed that t's posterior probability exceeds the one of t' (which might even be zero in case of a refuted theory). Nonetheless, although the predictions of t' are fraught with a higher risk of becoming falsified, t' might be preferable because of being highly informative. This reasoning applies even when t' is known to be false: even if Newtonian mechanics are falsified by e, they seem preferable to another nearly tautologous theory. We thus get the following constraint, where $\alpha(t, e)$ is a function representing the epistemic value of a theory t in the light of given evidence e[11]:

(a_1) If $e \models t \wedge \neg t'$, it may be that $\alpha(t, e) < \alpha(t', e)$.

Now assume that t and t' are both either entailed by e or refuted by e. In the former case, we are well-advised to opt for the more informative theory while this should not generally be the case among falsified theories. Otherwise, one could easily ameliorate the value of a falsified theory by adding any arbitrary prediction whatsoever.[12] For example, let t be Newtonian mechanics and t' the conjunction of t and one of Nostradamus' prophecies, then t' does not seem preferable to t. Thus, we add the following two constraints:

(a_2) If $e \models t \wedge t'$ and $t \models t'$, then $\alpha(t, e) \geq \alpha(t', e)$.
(a_3) If $e \models \neg t \wedge \neg t'$ and $t \models t'$, then it may be that $\alpha(t, e) < \alpha(t', e)$.

The next constraint is concerned with cases in which two rival theories explain the relevant evidence e.[13] In this case, if t is more demanding in the sense that $t \models t'$, then we should go for the weaker theory t'. The reason is that if this weaker theory allows us to predict e without any further assumptions, then *in the light of* e (and only e), t' is the better theory.

(a_4) If $t \models t'$ and $t' \models e$, then $\alpha(t, e) \leq \alpha(t', e)$.

Furthermore, predictive success should be reflected in the measure's value. That is, if t predicts e', then this should somehow be reflected in t's scientific value.

[11] (a_1)–(a_3) are standard verisimilitude-assumptions (cf. Cevolani 2011; Niiniluoto 1987, pp. 232ff.). (a_4)–(a_5) are taken from Zamora-Bonilla 1996; variants of (a_6) can be found in Glass 2007 and Kuipers 2009.

[12] This is similar to the famous *child's play objection* against content-definitions of verisimilitude (cf. Tichý 1974).

[13] Both in the case of explanation and in the case of predictive success to be considered below we assume simple cases of deductive explanation/success. This is not to say that these are the only relevant such cases; however, in the given context the focus is exclusively on these salient cases.

Table 1 Constraint-based evaluation of coherence measures

Measure	(a_1)	(a_2)	(a_3)	(a_4)	(a_5)	(a_6)
$\mathcal{D}$	–	✓	–	–	✓	✓
$\mathcal{O}$	–	✓	–	✓	✓	✓
$\mathcal{S}_d$	–	✓	–	✓	✓	–
$\mathcal{S}_k$	–	✓	–	–	✓	✓
$\mathcal{S}_f$	–	✓	–	✓	✓	✓

(a_5) If $t \models e'$, then $\alpha(t, e) \leq \alpha(t, e \wedge e')$.

The last constraint to be considered is concerned with a pair of theories t, t' such that t' has greater probability and also greater likelihood on e. The constraint reads as follows:

(a_6) If $\Pr(t|e) \leq \Pr(t'|e)$ and $\Pr(e|t) \leq \Pr(e|t')$, then $\alpha(t, e) \leq \alpha(t', e)$.

A summary of the results for all considered coherence measures is given in Table 1.[14]

The upshot is that *no* coherence measure satisfies all constraints. More specifically, all measures do at least violate (a_1) and (a_3). This means that none of these measures adequately captures the intuition that falsified theories can be more or less adequate in the light of falsifying data.

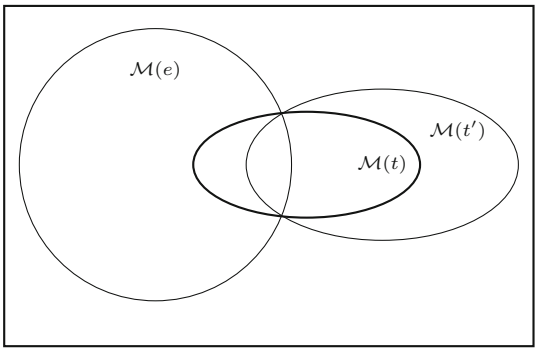
Another conclusion that can be drawn is that, although none of the measures satisfies all constraints, $\mathcal{O}$ and $\mathcal{C}_f$ outperform the other measures in that they satisfy all other constraints. $\mathcal{O}$ and $\mathcal{C}_f$ additionally account for the fact that if e is "at least as close to t than to t'", then the degree of coherence of (t, e) also exceeds the one of (t', e). The notion of closeness underlying this constraint can be illustrated as follows, where for each proposition φ, $\mathcal{M}(\varphi)$ is the set of *models* satisfying φ (cf. Kuipers 1992):

Here, the disagreement between t and e, $\mathcal{M}(t)\Delta\mathcal{M}(e)$, is a proper subset of the disagreement between t' and e, $\mathcal{M}(t')\Delta\mathcal{M}(e)$.[15] Therefore it seems that in light of e, t is preferable to t'. It can easily be shown that indeed both coherence measures also account for this intuition (cf. Zamora-Bonilla 1996).

In what follows I will argue that the coherence measures' deviation regarding the constraints pertaining to the adequacy of falsified theories is due to a too simplistic idea of applying coherence measures. More precisely, I will introduce a model of *two-sided coherence* that allows to repair the deficiencies of these measures in an intuitively satisfying way.

[14]The proofs only require elementary probability theory and straightforward arithmetic manipulations. For the proofs pertaining to measure $\mathcal{O}$ see Zamora-Bonilla 1996. There, $\Pr(e)^{-1} \times \mathcal{O}(t, e)$ is proposed as a probabilistic measure of verisimilitude. For further discussions see also Zamora-Bonilla 2002, 2013.

[15]Let Γ and Σ be two sets, then the symmetric difference, $\Gamma\Delta\Sigma$, is defined as follows: $\Gamma\Delta\Sigma = (\Gamma \setminus \Sigma) \cup (\Gamma \setminus \Sigma)$.



4 Two-Sided Coherence Measures

What causes the problems with coherence measures when it comes to assessing the value of falsified theories is their dependence on the probability of the conjunction of the hypothesis and the evidence under consideration. Of course, if the hypothesis and the evidence are logically incompatible, then this probability will always be equal to 0. Hence, in what follows I will suggest to split each theory t into a set t_e^+ of claims that are entailed by the evidence, a set t_e^- of claims that are refuted by e, and the remainder set t_e^r containing those claims that are neither entailed nor refuted. In the context of propositional languages, which will be our focus in the last sections, one can think of a *claim* c of a theory t as a disjunction $\varphi = \pm p_{i_1} \vee \ldots \vee \pm p_{i_j}$ of different and numerically ordered literals ($1 \leq i_1 < \ldots < i_j \leq n$) such that $t \models \varphi$ and there is no other such disjunction ψ such that $t \models \psi$ and ψ is logically stronger than φ.[16] The *set* of claims of a theory t will accordingly be denoted by $\mathbb{C}_t$. To illustrate, assume that $t = p_1 \wedge p_2 \wedge p_3$, $t' = (p_1 \vee p_2) \wedge (p_1 \vee \neg p_2)$ and $e = p_1 \wedge \neg p_2$, then the set of claims of t is $\mathbb{C}_t = \{p_1, p_2, p_3\}$. Hence, $t_e^+ = p_1$, $t_e^- = p_2$ and $t_e^r = p_3$. Accordingly, $\mathbb{C}_{t'} = \{p_1\} = t'^+_e$ and $t'^-_e = t'^r_e = \emptyset$.

Now two-sided coherence measures are based on the following intuition: let $e = \varphi_1 \wedge \ldots \wedge \varphi_n$ be a piece of evidence (where each φ_i is a disjunction of literals),[17] then $\overline{e} = \neg\varphi_1 \wedge \ldots \wedge \neg\varphi_n$ is called the *inverse* of e (cf. Zwart 2001). The idea of two-sided coherence is something like a trade-off between the following basic intuitions:

(i) the best theory coheres perfectly with e
(ii) the worst theory coheres perfectly with $\overline{e}$

[16] Cf. Schurz and Weingartner 2010 for similar characterizations.

[17] It is well-known that in propositional logics each formula can be translated into an equivalent formula in conjunctive normal form.

Two straightforward ideas for how to spell out the intuition of a trade-off between these two aspects are given by the following recipes. Within these formulae, "*Coh*" denotes an arbitrary probabilistic measure of coherence.

$$Coh^d(t, e, \overline{e}) = Coh(\bigwedge t_e^+, e) - Coh(\bigwedge t_e^-, \overline{e})$$
$$Coh^r(t, e, \overline{e}) = \log_2 \left[Coh(\bigwedge t_e^+, e) \big/ Coh(\bigwedge t_e^-, \overline{e}) \right]$$

Thus, a two-sided coherence measure is strictly increasing the more its verified part t_e^+ coheres with e, ceteris paribus, and strictly decreasing the more the part t_e^-, that is incompatible with e, coheres with the inverse $\overline{e}$. One can see immediately that two-sided coherence measures satisfy (a_1) and (a_3): given a tautologous theory t and a highly specific and almost true theory t', $Coh(\bigwedge t_e^+, e)$ will usually be small while $Coh(\bigwedge t'^+_e, e)$ can be sufficiently high in order to compensate for the minuend $Coh(\bigwedge t'^-_e, \overline{e})$. Similarly, falsified theories might differ largely in their verified and falsified set of claims and, a fortiori, differ in their degree of coherence as measured by a two-sided coherence measure.

5 Conjunctive Theories and Expected Coherence

This section is devoted to an assessment of two-sided coherence measures. In order to demonstrate their utility in the context of theory assessment, I will show that a specific two-sided coherence measure, the one based on the overlap-measure of coherence $\mathcal{O}$, establishes a link between coherence and truthlikeness: focusing on *conjunctive theories* it will be shown that a theory that coheres better with the available data is also assigned a higher degree of *estimated verisimilitude*.[18]

5.1 Estimated Verisimilitude for Conjunctive Theories

In a number of recent publications, Cevolani, Crupi and Festa developed a "basic feature approach" to verisimilitude.[19] The basic idea underlying this approach can be illustrated within the propositional language $\mathcal{L}_n$ with n atomic propositions $p_1, \ldots, p_n$ as follows: call $c = \pm p_1 \wedge \ldots \wedge \pm p_n$ a *constituent*, then c represents a possible state of affairs as described in $\mathcal{L}_n$. Assume furthermore that $\mathcal{C}$ denotes the set of all such constituents, then there is one and only one constituent $c^\star$ that delivers the whole truth about the domain under consideration. A *conjunctive theory* ("c-theory") t of $\mathcal{L}_n$ is a conjunction of k literals, i.e. each such c-theory has the following form:

$$\pm p_{i_1} \wedge \ldots \wedge \pm p_{i_k}$$

[18]The reason for focusing on this measure is that it performed well in the evaluation in Table 1. An extensive survey of different approaches is beyond the scope of the present paper.

[19]Cf. Cevolani et al. (2010, 2011) and references therein.

Now let $\mathcal{B} = \{p_1, \neg p_1, \ldots, p_n, \neg p_n\}$ denote the set of all literals of $\mathcal{L}_n$, then for each c-theory t define the set $\mathcal{B}_t = \{\varphi | \, t \models \varphi\} \cap \mathcal{B}$ of *basic claims* of t. It can easily be shown that in the context of c-theories, $\mathcal{B}_t = \mathbb{C}_t$. Now given the true constituent $c^\star$, each c-theory is decomposable into the true part $\mathcal{B}_t^+ = \{\varphi \in \mathcal{B}_t | \, c^\star \models \varphi\}$ and the false part $\mathcal{B}_t^- = \{\varphi \in \mathcal{B}_t | \, c^\star \models \neg\varphi\}$. Accordingly, Cevolani et al. (2011) define the *degree of true content* of t, $T{-}cont(t, c^\star)$, and the *degree of false content* of t, $F{-}cont(t, c^\star)$, as follows:

$$T{-}cont(t, c^\star) = |\mathcal{B}_t^+| \big/ n \qquad F{-}cont(t, c^\star) = |\mathcal{B}_t^-| \big/ n$$

The degree of verisimilitude of a theory t in light of the true constituent $c^\star$, $\mathrm{Vs}(t, c^\star)$, can now be calculated as follows (cf. Cevolani et al. 2011):

$$\mathrm{Vs}(t, c^\star) = T{-}cont(t, c^\star) - F{-}cont(t, c^\star)$$

It turns out that as far as c-theories are concerned, Vs is a special case among a large number of existing proposals to measuring verisimilitude.[20] It thus establishes a touchstone of verisimilitude intuitions that can be utilized in order to assess two-sided coherence measures.

However, the cases that have been considered so far presuppose that there is no uncertainty as regards the true state of the world. In natural environments this is almost never the case though. Accordingly, the *degree of estimated verisimilitude* in light of given evidence e is the expected degree of verisimilitude for all constituents weighted by their posterior probabilities:

$$\mathrm{EVs}(t, e) = \sum_{c \in \mathcal{C}} \mathrm{Vs}(t, c) \times \Pr(c|e)$$

One can easily check that for conclusive evidence such that $\Pr(c|e) = 1$ for constituent $c = c^\star$, the degree of estimated verisimilitude equals the degree of verisimilitude as measured by Vs.

5.2 Making the Link: Expected Coherence and Estimated Verisimilitude

So far, we concentrated on the degree of coherence between a theory and the available evidence. However, when assessing the degree of estimated verisimilitude, scholars usually calculate the expected value of verisimilitude for each theory in light of the given constituents. We can do the same for coherence. The idea now

[20] Among these are the ones proposed by Kuipers (1982), Oddie (1986), Schurz and Weingartner (1987, 2010), Brink and Heideman (1987) and Gemes (2007). Cf. Cevolani et al. 2011.

is to assess the value of a scientific theory by calculating the expectation value of coherence as a weighted average of the degrees of coherence between t on the one hand and the set of constituents on the other. This can be accomplished as follows:

$$\mathbb{E}(Coh(t,e)) = \sum_{c \in \mathbb{C}} \Pr(c|e) \times Coh(t, c, \overline{c})$$

Of course the question is what probability distribution is assumed for the calculation of the various degrees of coherence between t and each constituent c. In what follows I will assume a flat prior distribution $\Pr^*$ over $\mathcal{C}$, i.e. a probability distribution that assigns equal marginal probabilities to all constituents c of the language.[21]

Now, to illustrate the utility of two-sided coherence measures, we feed one of our favorite coherence measures, the overlap-based measure $\mathcal{O}$, into the ratio-based recipe Coh^r. For this measure we can prove the following result[22]:

Theorem 5.1. *Let t, t' be two c-theories and e the available evidence such that $\mathbb{E}(\mathcal{O}^r(t,e)) > \mathbb{E}(\mathcal{O}^r(t',e))$, then* $\mathrm{EVs}(t,e) > \mathrm{EVs}(t',e)$.

Thus, a higher degree of coherence between a theory and the given (probabilistic) evidence entails a higher degree of expected verisimilitude for this theory. Accordingly, we seem well advised to stick to theories that cohere with the available evidence.

6 Conclusion

This paper examined the virtue of a coherentist approach to assessing scientific theories. It was shown that although existing proposals to measuring coherence satisfy a number of adequacy constraints, none of them satisfies constraints pertaining to a comparison involving falsified theories. Consequently, two-sided coherence measures were introduced that allow for reasonable assessments of falsified theories as well. In addition, the paper argued for the utility of two-sided coherence measures by establishing a link between coherence orderings generated by a specific two-sided coherence measure and the corresponding orderings of estimated truthlikeness. These results support the positive findings regarding the coherentist approach to theory comparison by Brössel (2013).

[21] These *logical* probability distributions have famously been discussed by Carnap (1962).

[22] The same holds if we instead choose the firmness-based measure $\mathcal{S}_f$. However, a detailed investigation of the other measures is beyond the scope of the present paper. The proof of Theorem 5.1 is given in the appendix.

Appendix: A Proof of Theorem 5.1

Let $e_{\Pr}$ be a distribution of probabilities $\Pr(c|e)$ for each constituent $c \in \mathcal{C}$. The degree of coherence between c and an arbitrary c-theory $t = p_{i_1} \wedge \ldots \wedge p_{i_k}$ is given by the following formula:

$$\mathcal{O}(t, c, \overline{c}) = \log_2 \left[\mathcal{O}(\bigwedge t_c^+, c) \big/ \mathcal{O}(\bigwedge t_c^-, \overline{c}) \right] \tag{1}$$

It can easily be shown that the following identity holds:

$$\mathcal{O}(t, c, \overline{c}) = \log_2 \left[\frac{1}{\Pr(\bigwedge t_c^+ | c)} + \frac{1}{\Pr(c | \bigwedge t_c^+)} - 1 \right]^{-1} \tag{2}$$

$$- \log_2 \left[\frac{1}{\Pr(\bigwedge t_c^- | \overline{c})} + \frac{1}{\Pr(\overline{c} | \bigwedge t_c^-)} - 1 \right]^{-1} \tag{3}$$

Now for conjunctive theories, both $\Pr(\bigwedge t_c^+ | c)$ and $\Pr(\bigwedge t_c^- | \overline{c})$ are equal to 1. Hence, Eq. 2 reduces to the following formula:

$$\mathcal{O}(t, c, \overline{c}) = \log_2 \Pr(c | \bigwedge t_c^+) - \log_2 \Pr(\overline{c} | \bigwedge t_c^-) \tag{4}$$

Exploiting Bayes' theorem, we get (remember that Pr is the logical probability $\Pr^*$):

$$\mathcal{O}(t, c, \overline{c}) = \log_2 \frac{\Pr(c)}{\Pr(\bigwedge t_c^+)} - \log_2 \frac{\Pr(\overline{c})}{\Pr(\bigwedge t_c^-)} = |t_c^+| - |t_c^-| \tag{5}$$

Hence, $\mathcal{O}(t, c, \overline{c})$ is strictly increasing in $|t_c^+|$ and strictly decreasing in $|t_c^-|$ for each constituent c. The same obviously holds for $\mathrm{Vs}(t, c)$. Therefore, for each pair of c-theories t, t' and each constituent the following equivalence holds:

$$\mathcal{O}(t, c, \overline{c}) \geq \mathcal{O}(t', c, \overline{c}) \quad \Leftrightarrow \quad \mathrm{Vs}(t, c) \geq \mathrm{Vs}(t', c) \tag{6}$$

Hence, (5), (6) and the definition of Vs together entail that

$$\sum_{c \in \mathbb{C}} \Pr(c|e) \times \mathcal{O}(t, c, \overline{c}) \geq \sum_{c \in \mathbb{C}} \Pr(c|e) \times \mathcal{O}(t', c, \overline{c})$$

$$\Leftrightarrow$$

$$\sum_{c \in \mathbb{C}} \Pr(c|e) \times \mathrm{Vs}(t, c) \geq \sum_{c \in \mathbb{C}} \Pr(c|e) \times \mathrm{Vs}(t', c)$$

Thus, we get the desired result that if $\mathbb{E}(\mathcal{O}(t, e)) > \mathbb{E}(\mathcal{O}(t, e))$, then $\mathrm{EVs}(t, e) > \mathrm{EVs}(t', e)$ (and also vice versa). □

References

Bar-Hillel, Y., & Carnap, R. (1953). Semantic information. *British Journal for the Philosophy of Science, 4*, 147–157.

Bonjour, L. (1985). *The structure of empirical knowledge*. Cambridge/London: Harvard University Press.

Brink, C., & Heideman, J. (1987). A verisimilar ordering of theories phrased in a propositional language. *British Journal for the Philosophy of Science, 38*, 533–549.

Brössel, P. (2013). Assessing theories: The coherentist approach. *Erkenntnis, 79*, 593–623.

Carnap, R. (1962). *The logical foundations of probability* (2nd ed.). Chicago: Chicago University Press.

Cevolani, G. (2011). Verisimilitude and strongly semantic information. *Ethics & Politics, 2*, 159–179.

Cevolani, G., & Tambolo, L. (2013). Progress as approximation to the truth: A defence of the verisimilitudinarian approach. *Erkenntnis, 78*, 921–935.

Cevolani, G., Crupi, V., & Festa, R. (2010). The whole truth about Linda: Probability, verisimilitude and a paradox of conjunction. In M. D'Agostino et al. (Eds.), *New essays in logic and philosophy of science* (pp. 603–615). London: College Publications.

Cevolani, G., Crupi, V., & Festa, R. (2011). Verisimilitude and belief change for conjunctive theories. *Erkenntnis, 75*, 183–202.

Crupi, V. (2015) Confirmation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2015 Edition). http://plato.stanford.edu/archives/sum2015/entries/confirmation/

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science, 74*, 229–252.

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

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

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

Fitelson, B. (2004). Two technical corrections to my coherence measure. http://www.fitelson.org/coherence2.pdf.

Floridi, L. (2004). Outline of a theory of strongly semantic information. *Minds and Machines, 14*, 197–221.

Gemes, K. (2007). Verisimilitude and content. *Synthese, 154*, 293–306.

Glass, D. H. (2002). Coherence, explanation, and Bayesian networks. In M. ONeill et al. (Eds.), *Artificial intelligence and cognitive science* (pp. 177–182). Berlin/Heidelberg: Springer.

Glass, D. H. (2007). Coherence measures and inference to the best explanation. *Synthese, 157*, 275–296.

Good, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation, 19*, 294–299.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge: Cambridge University Press.

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

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

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

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

Kuipers, T. A. F. (1982). Approaching descriptive and theoretical truth. Erkenntnis, 18, 343–378.

Kuipers, T. A. F. (1987). A structuralist approach to truthlikeness. In T. A. F. Kuipers (Ed.), *What is closer-to-the-truth?* (pp. 79–99). Amsterdam: Rodopi.

Kuipers, T. A. F. (1992). Naive and refined truth approximation. *Synthese, 93*, 299–342.

Kuipers, T. A. F. (2000). *From instrumentalism to constructive realism. On some relations between confirmation, empirical progress, and truth approximation*. Dordrecht: Kluwer.

Kuipers, T. (2009). Empirical progress and truth approximation by the 'Hypothetico-Probabilistic Method'. *Erkenntnis, 70*, 313–330.
Lehrer, K. (1974). *Knowledge*. New York: Oxford University Press.
Meijs, W. (2006). Coherence as generalized logical equivalence. *Erkenntnis, 64*, 231–252.
Miller, D. (1974). Popper's qualitative theory of verisimilitude. *The British Journal for the Philosophy of Science, 25*, 166–177.
Niiniluoto, I. (1984). *Is science progressive?*. Dordrecht: Reidel.
Niiniluoto, I. (1987). *Truthlikeness*. Dordrecht: Reidel.
Niiniluoto, I. (1999). *Critical scientific realism*. Oxford: Oxford University Press.
Niiniluoto, I. (2011). Scientific progress. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2011 ed.). http://plato.stanford.edu/archives/sum2011/entries/scientific-progress. Accessed 04 Mar 2014.
Oddie, G. (1986). *Likeness to truth*. Dordrecht: Reidel.
Oddie, G. (2008). Truthlikeness. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2014 ed.). http://plato.stanford.edu/archives/spr2014/entries/truthlikeness. Accessed 05 Mar 2014.
Olsson, E. J. (2002). What is the problem of coherence and truth? *The Journal of Philosophy, 99*, 246–272.
Popper, K. R. (1963). *Conjectures and refutations*. London: Routledge and Kegan Paul.
Popper, K. R. (1968). *The logic of scientific discovery*. London: Hutchinson.
Popper, K. R. (1972). *Objective knowledge. An evolutionary approach*. Oxford: Clarendon Press.
Roche, W. (2013). Coherence and probability. A probabilistic account of coherence. In M. Araszkiewicz & J. Savelka (Eds.), *Coherence: Insights from philosophy, jurisprudence and artificial intelligence* (pp. 59–91). Dordrecht: Springer.
Schupbach, J. N. (2011). New hope for Shogenji's coherence measure. *The British Journal for the Philosophy of Science, 62*, 125–142.
Schurz, G., & Weingartner, P. (1987). Verisimilitude defined by relevant consequence elements. In T. Kuipers (Ed.), *What is closer-to-the-truth?* (pp. 47–77). Amsterdam: Rodopi.
Schurz, G., & Weingartner, P. (2010). Zwart and Franssen's impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude. *Synthese, 172*, 415–436.
Shogenji, T. (1999). Is coherence truth conducive? *Analysis, 59*, 338–345.
Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435–467.
Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science, 22*, 1–24.
Tichý, P. (1974). On Popper's definition of verisimilitude. *The British Journal for the Philosophy of Science, 25*, 155–160.
Zalabardo, J. (2009). An argument for the likelihood-ratio measure of confirmation. *Analysis, 69*, 630–635.
Zamora-Bonilla, J. P. (1996). Verisimilitude, structuralism and scientific progress. *Erkenntnis, 44*, 25–48.
Zamora-Bonilla, J. P. (2002). Verisimilitude and the dynamics of scientific research programmes. *Journal for General Philosophy of Science, 33*, 349–368.
Zamora-Bonilla, J. P. (2013). Why are good theories good? Reflections on epistemic values, confirmation, and formal epistemology. *Synthese, 190*, 1533–1553.
Zwart, S. D. (2001). *Refined verisimilitude*. Dordrecht: Kluwer.

A Verisimilitudinarian Rebuttal of a Recent Attack Against Realism

Luca Tambolo

1 Introduction

Karl Popper (1963, 1972) was the first to put forward a formal explication of the view that scientific progress can be accounted for in terms of the increasing verisimilitude—or equivalently, truthlikeness, or approximation to the truth—of our theories. After Miller (1974) and Tichý (1974) independently demonstrated that, on the basis of Popper's explication of the concept of verisimilitude, a false theory can never be closer to the truth than another (true or false) theory, various philosophers proposed—partially conflicting with each other—post-Popperian theories of verisimilitude, based on explications of such concept that avoid the flaws of Popper's account (see, e.g., Oddie 1986; Kuipers 1987, 2000; Niiniluoto 1987, 1999; Schurz and Weingartner 1987, 2010; and see Niiniluoto 1998 and Oddie 2014 for surveys of the history of such theories).

During the 1980s, the technical work on the notion of verisimilitude by the above mentioned authors gave rise to a fallibilist, and yet robustly realist, post-Popperian approach to scientific progress mainly championed by Ilkka Niiniluoto and Theo Kuipers, which has sometimes been termed "verisimilitudinarian" (henceforth: VS) in the literature (Festa 2007; Cevolani and Tambolo 2013a, b).

Within the context of the most recent debate on the nature of scientific progress, various authors have attacked VS (Bird 2007, 2008; Rowbottom 2008, 2010; see Cevolani and Tambolo 2013a and Niiniluoto 2014 for a defense of VS). In this paper, we aim at contributing to the debate by rebutting one of these attacks, due to Carlotta Piscopo and Mauro Birattari (2010). According to them, verisimilitude does not

L. Tambolo (✉)
Department of Humanistic Studies, University of Trieste, c/o via Casona, 7, 40043 Marzabotto, Italy
e-mail: ltambolo@gmail.com

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_2

provide the defense of scientific realism that its proponents strive for, due to the fact that it does not satisfactorily perform the double role—*constitutive* and *regulative*—that it is supposed to play within VS. More precisely, Piscopo and Birattari argue that, although the idea of approximation to the truth may be useful when deployed in a regulative role, i.e., as a motivation, or inspiration, in putting forward theories, verisimilitude does not provide objective grounds for choosing, among competing theories, the one which is closer to the truth, and therefore constitutes a case of progress; consequently, they maintain, verisimilitude fails when deployed in what they call a constitutive role. Verisimilitude, they conclude, ought to be excluded from theory appraisal and confined to the context of theory construction. In what follows, after briefly recalling some basic features of VS (Sect. 2), we shall scrutinize the constitutive/regulative distinction on which Piscopo and Birattari's criticism of VS is based (Sect. 3). We shall then argue that, contrary to what they claim, it is not the case that, within VS, there is one notion—verisimilitude—playing a double role: rather, there are two different notions—*verisimilitude* and *estimated verisimilitude*—playing two different roles. In Sect. 4, we shall rebut their criticism by showing that: (*a*) it stems from the neglect of the distinction between verisimilitude and estimated verisimilitude, and of the related, crucial distinction between *real progress* and *estimated progress*; (*b*) it presupposes the adoption of a standard for the justification of the choice among theories which is too demanding—so demanding that, as it turns out, it requires the embrace of an infallibilist view of science. In Sect. 5, we shall offer some brief concluding remarks.

2 A Brief Overview of VS

VS revolves around the idea that the increasing verisimilitude, or the decreasing distance from the truth, of our theories is the key ingredient for progress. For instance, according to VS, such theoretical transitions as that from Newton's theory (T_1) to Einstein's theory (T_2) are to be viewed as progressive because, although T_2 is, strictly speaking, presumably false, we have what seem to be good reasons to believe that T_2 is closer to the truth than T_1. The key tenets of VS can then be briefly formulated as follows:

1. since some false theories are closer to the truth than others, a false but highly verisimilar theory can constitute a genuine approximation to the truth, and hence an instance of progress, when it is adopted to replace a less verisimilar theory;
2. we can devise methods to fallibly assess, given the available evidence, which among two competing theories is closer to the truth.

In other words, within VS truth is taken to be the ideal goal of inquiry. However, the proponents of VS readily acknowledge that even our most successful theories are typically, strictly speaking, false, if only because they involve the use of various idealizing assumptions. In view of this, within VS, the main cognitive aim of

scientific inquiry is assumed to be the search for highly verisimilar theories, i.e., theories which, although presumably false, are close to the truth.

For the purposes of the present discussion, the idea of verisimilitude can be characterized, in informal terms, by saying that a theory T is highly verisimilar (or equivalently, truthlike, or close to the truth) if it says many things about a target domain, as described in a given language, and if many of these things are (almost exactly) true. Thus, the (degree of) verisimilitude of a theory depends on both its content, i.e., how much the theory says, and its accuracy, i.e., how much of what the theory says is in fact true. In other words, high verisimilitude requires an appropriate combination of truth and content. In spite of the problems encountered by Popper's explication of verisimilitude, this is exactly the insight underlying his approach to the issue, since he claimed that verisimilitude "represents the idea of *approaching comprehensive truth*" (1963, p. 237, emphasis added).[1]

Within VS, a distinction dating back to Popper (1963, Ch. 10) is drawn between the logical problem of verisimilitude and the epistemic problem of verisimilitude. Addressing the *logical problem* means providing an appropriate definition of the notion of *verisimilitude*, such that a comparison between any two theories T_1 and T_2 with respect to their closeness to the truth, assumed to be known, is possible. Addressing the *epistemic problem* means providing an appropriate definition of the notion of *estimated verisimilitude*, such that given a certain body of evidence, it is possible to rationally conjecture, for instance, that a theory T_1 is close to the truth, supposed to be unknown, or that T_2 is closer to the truth than T_1. In other words, the verisimilitude of a theory T, which is also called its "real verisimilitude," or "absolute verisimilitude," is a measure of how close T is to the truth, i.e., the ideal goal of inquiry; while the estimated verisimilitude of T constitutes our fallible assessment of its real, or absolute, verisimilitude.

Since in the most interesting cases truth is unfortunately unknown, the epistemic problem of verisimilitude has been a central concern for both the main proponents of VS. Due to space limitations, here we shall not enter into the details of the solutions that Niiniluoto and Kuipers provide to such problem: it will suffice to mention that Niiniluoto puts forward a quantitative notion of expected verisimilitude (1987, 1999, 2014), while Kuipers prefers a non-quantitative approach to the estimation of verisimilitude (1987, 2000, 2014). In both cases, the rules for theory-choice that they defend are ways of making rational conjectures, based on the available evidence, concerning a theory's closeness to the truth. In other words, the notion of estimated

[1]The nature of the combination of truth and content mentioned in the text can perhaps be more easily appreciated by considering the difference between the notions of approximate truth and verisimilitude. A theory T is approximately true, or accurate, if T is close to being true, i.e., roughly, if many of the things that T says are true; T is highly verisimilar if, as recalled above, T is close to the (whole) truth about a certain domain, i.e., if T says many things—T is informative—and many of these things are (almost exactly) true. In view of the fact that approximate truth is only one of the ingredients of verisimilitude, a theory with a high degree of approximate truth, which however does not convey much information, may well have a low degree of verisimilitude (see Niiniluoto 1987, pp. 176–177 and 217 ff.).

verisimilitude allows one to say that, given a certain body of evidence, a theory T_2 *seems* more verisimilar than a theory T_1, i.e., that it is reasonable to claim that T_2 is closer to the truth than T_1.

Finally, from the distinction between the logical and the epistemic problem of verisimilitude, it follows that progress can be characterized, within VS, not only as *real progress*, construed as increasing (real, or absolute) verisimilitude, but also as *estimated progress*, construed as increasing estimated verisimilitude. In other words, within VS the transition from the embrace of a theory T_1 to the embrace of a theory T_2 is considered as progressive, in the sense of being a case of real progress, if and only if T_2 is actually closer to the truth than T_1; while it is said to *seem* progressive, in the sense of being a case of estimated progress, if and only if, given the available evidence, T_2 is estimated to be closer to the truth than T_1.

The above sketch of the main features of VS will serve as a starting point for rebutting, in Sect. 4, Piscopo and Birattari's criticism against it.

3 Two Roles of Verisimilitude Within VS?

As mentioned in Sect. 1, Piscopo and Birattari (2010) distinguish between two roles of the notion of verisimilitude within VS: a regulative one and a constitutive one. At the very beginning of their paper, they claim that "the dissatisfaction with the notion of truthlikeness" as a tool used to defend a realist account of scientific progress "concerns the double role that this notion plays" (ibid., p. 379) within VS.[2] They then go on to argue that, in view of the difficulties encountered by the notion of verisimilitude as deployed in a constitutive role (more on this below), a reasonable solution for the proponents of VS would be that of attributing to verisimilitude "only a regulative role" (ibid., p. 385). As we shall show, one of the problems with this line of argument lies in the fact that it involves a mischaracterization of VS; however, let us first examine their distinction between a regulative and a constitutive role, and the reasons why their criticism of verisimilitude as deployed in a constitutive role matters within the current debate on the nature of progress.

In the broadly Kantian terminology that Piscopo and Birattari use, a notion plays a *regulative role* "if it is assumed and/or used as an inspiration in putting forward a theory, but it does not play then a role in its assessment" (ibid., p. 381). Within VS, verisimilitude plays a regulative role since it performs the function of "motivating, inspiring and guiding scientific research" (ibid.). In other words, in the phase of theory construction, verisimilitude plays a regulative role in the sense that

[2]It should be mentioned here that Piscopo and Birattari explicitly argue only against the so-called "similarity approach" developed by Oddie (1986) and Niiniluoto (1987, 1999), and not against VS, or against verisimilitude-based approaches to theory-choice and progress in general. However, in light of our overview of the main features of VS (Sect. 2), it should be clear that their line of argument, as it will be presented in this section, is aimed at applying also to Kuipers' (2000) version of VS—and of course, the same applies to our rebuttal of their criticism.

"before a newly conceived theory is empirically evaluated, a scientist anticipates that it is, in some sense, closer to the truth than the current one" (ibid., p. 382). Piscopo and Birattari have no complaints concerning this latter use of verisimilitude (see ibid., pp. 384–386), and they insist that the aim of Popper's work on such notion was precisely to defend approximation to the truth as a regulative ideal of scientific inquiry.[3] Their objections are targeted at the use of verisimilitude in a *constitutive role*. In their terminology, a notion is used in such a role if "it is adopted as an ultimate criterion to select one out of two rival theories" (ibid., p. 381). Verisimilitude, they claim, "clearly plays a *constitutive* role" within VS, "and within the realist epistemology more in general," since according to the proponents of VS, "the conclusive criterion for preferring a theory T_2 to a theory T_1 is that T_2 better corresponds to reality" (ibid.).

It is precisely with regards to its role in theory-choice that, Piscopo and Birattari argue, verisimilitude turns out to be an inadequate tool for the defense of a realist account of progress. In fact, according to them, verisimilitude "does not fulfill the requirement of objectivity" (ibid., p. 384) that it should fulfill in order for realists to be able to support their claims concerning the epistemological significance that one has to attach to the success of theories.

Such significance is, notoriously, a matter of eternal dispute between realists and antirealists. Realists typically claim that the success enjoyed by scientific inquiry would be an inexplicable fact, if scientific theories were not (at least approximately) true descriptions of (some of the relevant features of) the world. In this regard, Hilary Putnam defended the so-called "No Miracles Argument" by claiming that realism is "the only philosophy that doesn't make the success of science a miracle" (1975, p. 73), and others have proposed a verisimilitudinarian version of the No Miracles Argument, according to which one can draw—fallible and tentative—inferences from the success of a theory to its closeness to the truth (see, e.g., Niiniluoto 1999; Cevolani and Tambolo 2013b). Antirealists, on the other hand, deny the existence of any success-to-(approximate)truth/verisimilitude link, and typically point to the history of science as disproving the realists' claim that one can walk on what Larry Laudan called the "Upward Path" (1981, pp. 32–36). In fact, the history of science is replete with theories that, although embraced as successful for a certain period, later turned out to be false and were dropped by the scientific community; therefore, the antirealists argue, there is no reason to believe that our current theories fare any better than their predecessors in this regard. By the antirealists' lights, this so-called "Pessimistic Induction" shows that the claim that progress can be accounted for in terms of the increasing approximation to the truth of our theories is unsupported.

[3]For instance, in this regard Piscopo and Birattari quote Popper claiming that verisimilitude is a regulative notion, and "not an epistemological or an epistemic" (Popper 1963, p. 234) one, and that "we have no criterion of truth," so that we are only "guided by the idea of truth as a *regulative principle*" (ibid., p. 226).

Within this debate, Piscopo and Birattari side decidedly with antirealists, since they maintain that the notion of verisimilitude, deployed in a constitutive role, exhibits a "lack of objectivity" (2010, p. 385) which prevents it from providing an adequate response to the Pessimistic Induction (ibid., pp. 379–380). The passage in which they explain the reason for such conviction, with reference to Niiniluoto's version of VS, is worth quoting in full:

> The truth h_*, to which a scientific statement tends, is typically unknown and the only measure that can be computed is an estimated degree of truthlikeness: the *expected verisimilitude* $ver(g/e)$, which is a fallible indicator of the *degree of truthlikeness* $Tr(g, h_*)$. The expected verisimilitude $ver(g/e)$ indicates how close a statement g is to truth h_* on the basis of some empirical evidence e. The measure *ver* has therefore an empirical nature and is directly related to the success of the statement g. In this sense, *ver* allows one to select a theory and to evaluate its progressive character on the basis of empirical evidence. *Consequently, the claim that a theory is truthlike, as it is successful, can be disconfirmed by further evidence* (ibid., pp. 380–381, emphasis added).

In other words, their argument against VS runs as follows: (*a*) in order for verisimilitude to satisfactorily perform a constitutive role, estimates concerning the closeness to the truth of scientific statements and theories should not be revisable in the light of newly acquired evidence; (*b*) within VS, such estimates are acknowledged to be revisable in light of newly acquired evidence; (*c*) therefore, verisimilitude does not satisfactorily perform a constitutive role within VS, and the notion of verisimilitude should be restricted to the domain of theory construction, i.e., it should have only a regulative role. In Sect. 4, we shall rebut such criticism and suggest that it involves a mischaracterization of VS.

4 A Too Demanding Standard for Theory-Choice

The regulative/constitutive role distinction drawn by Piscopo and Birattari stems from a failure to appreciate that, with VS, it is not the case that there is one notion—verisimilitude—playing two roles. Rather, there are two notions—not coincidentally, each with its own name: verisimilitude and estimated verisimilitude—playing different roles. Curiously enough, in the passage quoted towards the end of Sect. 3, Piscopo and Birattari themselves mention Niiniluoto's distinction between expected verisimilitude and truthlikeness, and elsewhere they acknowledge that his "pivotal idea" (ibid., p. 380) consists in introducing an empirical estimate of the verisimilitude of theories, which works as a fallible indicator of their real, or absolute, verisimilitude. Nevertheless, in the course of their discussion, they repeatedly conflate the notions involved:

> [. . .] problems emerge in a constitutive context since the fact that truthlikeness is revisable challenges the central tenet of the realist epistemology: it eventually threatens the idea that what drives the actual selection of a theory, among rival ones, is its a priori objective truthlikeness and thus its better correspondence to reality (ibid., p. 383).

> As long as the truthlikeness of a theory plays a primary role in theory selection, it undergoes the empirical falsification and it challenges the key realist assumption that what drives the selection of a theory is its actual closeness to the truth (ibid., p. 384).

In both of the above passages the claim is made that, within VS, theory-choice is governed by the verisimilitude of theories. But this is not a claim that the proponents of VS make. As our discussion in Sect. 2 shows, it is one of the main tenets of VS that one can devise methods to fallibly *estimate*, given the available evidence, which among two (or more) competing theories is closer to the truth. Note that the point here is anything but purely terminological: it is not just that Piscopo and Birattari prefer to talk of the constitutive and regulative role of verisimilitude, while instead the proponents of VS like to speak of verisimilitude and estimated verisimilitude. For the latter distinction is motivated by the fact that, especially in the most interesting cases, there is no way to ascertain whether a certain belief or sets of beliefs exhibits a genuine correspondence to "the real world:" such correspondence is something to which we have no epistemic access. From this, it follows that a related distinction is made between *real progress*, on the one hand, and *estimated progress*, on the other hand.[4]

The failure to appreciate the above distinctions, of crucial importance within VS, in not inconsequential. As previously recalled, Piscopo and Birattari complain that the notion of verisimilitude, deployed in a constitutive role, ought to be relinquished, since it lacks the objectivity required to underpin the realists' claims concerning the link between success and (approximate)truth/verisimilitude, and consequently, the progressive character of the scientific enterprise. In this paper, we shall have nothing new to add to the existing literature on such a link. For the purposes of our present discussion, what matters is that the notion of objectivity that Piscopo and Birattari use in their criticism of VS sets a very high standard for the justification of theory-choice—indeed, a standard which is far too demanding.

In fact, their objection against theory-choice as characterized within VS boils down to the complaint that, being based on the available evidence, it is revisable in light of newly acquired evidence. In this regard, Piscopo and Birattari mention Laudan's (1981) seminal essay and remark that "the available historical evidence seems to indicate that the empirical success of theories does not *guarantee* either their genuine reference or their truthlikeness" (2010, p. 381, emphasis added). But the request for such a guarantee clearly amounts to a request for rules of theory-choice yielding non-revisable judgments concerning the closeness to the truth of theories; in other words, it amounts to a request for an *infallible* rule of theory-choice (on this, see Niiniluoto 2014).

[4]We should note in passing that, contrary to what one may tend to believe, there is only a partial overlap between the regulative/constitutive role distinction and the verisimilitude/estimated verisimilitude distinction. In fact, on the one hand, the notion of verisimilitude as deployed in a constitutive role seems to correspond quite well to the notion of estimated verisimilitude. On the other hand, the notion of verisimilitude as deployed in a regulative role, while conveying the idea of truth approximation as a guiding principle of scientific inquiry, is entirely silent concerning the task—central within VS—of precisely defining closeness to the truth.

As hinted in Sect. 2, nothing could be farther from the spirit of VS than the idea of an easily accessible, certified truth—an idea which can be entertained only by very naïve realists. Indeed, the only kind of realist who is likely to be looking for the infallible rule of theory-choice yearned for by Piscopo and Birattari is the one which John Worrall (2011, p. 159) has labeled "gung-ho realist," i.e., someone who claims that we must always believe in the truth of our currently accepted successful theories. In sum, as our discussion shows, Piscopo and Birattari criticize VS for being unable to provide something that—with good reasons—its proponents claim cannot be provided. At the beginning of their paper Piscopo and Birattari state that they aim at disentangling "why the notion of truthlikeness has been perceived as unsatisfactory to defend realism" (2010, p. 380). While we have no comments to offer on the reasons why VS has so far failed to win wider acceptance among philosophers of science, it seems to us that, if one of such reasons is the criticism discussed here, it is indeed a very poor reason.

5 Concluding Remarks

In this paper we defended VS against a criticism raised by Piscopo and Birattari (2010), who claim that verisimilitude does not provide objective grounds for theory selection, and therefore fails to support scientific realism in the way that the proponents of VS hope: it does not guarantee that the appraisal of competing theories will bring about the choice of the one which is closer to the truth. After briefly outlining the basic features of VS, and putting a special emphasis on the distinction, drawn by the supporters of VS, between real progress and estimated progress, we showed that Piscopo and Birattari's criticism, besides neglecting such crucial distinction, presupposes the embrace of an infallibilist view of science. In light of this—and of other rebuttals of recent attacks against VS (Cevolani and Tambolo 2013a, b; Niiniluoto 2014)—we conclude that VS cannot be discarded as easily as some of the contributors to the most recent debate on scientific progress seem to believe.

Acknowledgments Thanks are due to Gustavo Cevolani for many inspiring conversations on the material presented here and to two anonymous reviewers for their comments on a previous version of this paper. Financial support from PRIN grant *Models and Inferences in Science. Logical, Epistemological, and Cognitive Aspects* is gratefully acknowledged.

References

Bird, A. (2007). What is scientific progress? *Noûs, 41*, 64–89.

Bird, A. (2008). Scientific progress as accumulation of knowledge. A reply to Rowbottom. *Studies in History and Philosophy of Science, 39*, 279–281.

Cevolani, G., & Tambolo, L. (2013a). Progress as approximation to the truth. A defence of the verisimilitudinarian approach. *Erkenntnis, 78*, 921–935.
Cevolani, G., & Tambolo, L. (2013b). Truth may not explain predictive success, but truthlikeness does. *Studies in History and Philosophy of Science, 44*, 590–593.
Festa, R. (2007). Verisimilitude, cross classification, and prediction logic. Approaching the statistical truth by falsified qualitative theories. *Mind & Society, 6*, 91–114.
Kuipers, T. A. F. (1987). A structuralist approach to truthlikeness. In T. A. F. Kuipers (Ed.), *What is closer-to-the-truth?* (pp. 77–99). Amsterdam: Rodopi.
Kuipers, T. A. F. (2000). *From instrumentalism to constructive realism*. Dordrecht: Kluwer.
Kuipers, T. A. F. (2014). Empirical progress and nomic truth approximation revisited. *Studies in History and Philosophy of Science, 46*, 64–72.
Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science, 48*, 19–148.
Miller, D. (1974). Popper's qualitative theory of verisimilitude. *The British Journal for the Philosophy of Science, 25*, 166–177.
Niiniluoto, I. (1987). *Truthlikeness*. Dordrecht: Kluwer.
Niiniluoto, I. (1998). Verisimilitude: The third period. *The British Journal for the Philosophy of Science, 49*, 1–29.
Niiniluoto, I. (1999). *Critical scientific realism*. Oxford: Oxford University Press.
Niiniluoto, I. (2014). Scientific progress as increasing verisimilitude. *Studies in History and Philosophy of Science, 46*, 73–77.
Oddie, G. (1986). *Likeness to truth*. Dordrecht: Reidel.
Oddie, G. (2014). Truthlikeness. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Summer 2014 ed.).*http://plato.stanford.edu/archives/sum2014/entries/truthlikeness/
Piscopo, C., & Birattari, M. (2010). A critique of the constitutive role of truthlikeness in the similarity approach. *Erkenntnis, 72*, 379–386.
Popper, K. R. (1963). *Conjectures and refutations*. London: Routledge & Kegan Paul.
Popper, K. R. (1972). *Objective knowledge*. Oxford: Clarendon.
Putnam, H. (1975). *Mathematics, matter and method*. Cambridge: Cambridge University Press.
Rowbottom, D. P. (2008). N-rays and the semantic view of scientific progress. *Studies in History and Philosophy of Science, 39*, 277–278.
Rowbottom, D. P. (2010). What scientific progress is not: Against Bird's epistemic view. *International Studies in the Philosophy of Science, 24*, 241–255.
Schurz, G., & Weingartner, P. (1987). Verisimilitude defined by relevant consequence-elements. A new reconstruction of Popper's original idea. In T. A. F. Kuipers (Ed.), *What is closer-to-the-truth?* (pp. 47–78). Amsterdam: Rodopi.
Schurz, G., & Weingartner, P. (2010). Zwart and Franssen's impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude. *Synthese, 172*, 415–436.
Tichý, P. (1974). On Popper's definition of verisimilitude. *The British Journal for the Philosophy of Science, 25*, 155–160.
Worrall, J. (2011). Underdetermination, realism and empirical equivalence. *Synthese, 180*, 157–172.

Realistic Claims in Logical Empiricism

Matthias Neuber

1 Introduction

Logical Empiricism has for a long time been conceived of as a monolithic, one-dimensional, movement within early twentieth-century philosophy. As such, it frequently served as a contrast for later, opposing viewpoints, such as scientific realism, critical rationalism, or Kuhnian historical relativism. However, as more recent research has revealed, logical empiricism was much more multifaceted than commonly assumed. Especially the seminal contributions by Michael Friedman (1999), Friedrich Stadler (2001), and Thomas Uebel (2007) strongly indicate that the assumption of the existence of *varieties* of logical empiricism is clearly closer to the truth than the view of it as a narrow, quasi-dogmatic 'school.'

As concerns the debate over scientific realism, appraisals like the following, suggesting a strong *incompatibility* between scientific realism and logical empiricism, are still quite widespread:

> The philosophy of science in the twentieth century has been a battlefield between 'realist' and 'anti-realist' approaches. The interpretation of scientific theories, and the dispute about the cognitive significance of their theoretical terms and claims, provided a major impetus for the work of the Vienna Circle in the 1920s. The demise of logical positivism was followed by the rise of scientific realism within the analytic philosophy of science in the 1950s [. . .]. (Niiniluoto 1996, p. v)

There is little doubt that scientific realism became the dominant position in the philosophy of science in the second half of the twentieth century. And it cannot be denied that logical empiricism began to lose momentum. However, it must be

M. Neuber (✉)
Department of Philosophy, Universität Tübingen, Philosophisches Seminar, Bursagasse 1, Tübingen D-72070, Germany
e-mail: matthias.neuber@uni-tuebingen.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_3

seen that the logical empiricist program – at least in some of its guises – was not so different from the scientific realist account of science and nature. To be sure, there was a strong rejection of any form of *metaphysical* realism. But with respect to empirical science, most of the logical empiricists regarded realism not as a particularly contentious issue. Rather, they attempted to *argue for* the case of realism. More precisely, four kinds of arguments can be distinguished in this connection: a 'probabilistic argument,' a 'pragmatic argument,' an 'indispensability argument,' and an 'invariantist argument.' Given the variety of arguments, it is plausible to assume that logical empiricism and scientific realism are essentially *compatible with each other*.

It is the aim of the following considerations to clarify and fortify this compatibilist idea of 'realistic claims in logical empiricism.' I will attempt to make clear that the logical empiricists from the very beginning were rather open-minded toward an empirical, non-speculative, understanding of realism (Sect. 2). My goal is to demonstrate that this sort of programmatic open-mindedness developed into a (more or less) sophisticated commitment to the scientific realist agenda (Sect. 3) and that one specific articulation of 'realistic claims in logical empiricism,' namely the one delivered by Eino Kaila, comes close to current 'structural' realism (Sect. 4). By way of conclusion, it will be suggested that Kaila's (invariantist) approach gives rise to the establishment of an autonomous, measurement-based, account of structural realism (Sect. 5).

2 The Realism Issue: A Mere Pseudo-Problem?

To help clarify the idea of 'realistic claims in logical empiricism,' it is reasonable to begin with what might be called the 'received view' of the logical empiricist approach toward the realism issue. According to this received view, the realism issue is nothing but a *pseudo-problem*. And indeed: By examining the relevant writings of the relevant authors, one pretty soon discovers that the received view can easily be corroborated. Thus, for example, Rudolf Carnap, in his *Pseudoproblems in Philosophy* from 1928, explicitly states: "In the realism controversy, science can take neither an affirmative nor a negative position since the question has no meaning." (Carnap [1928a] 1968, p. 333) Quite similarly, Moritz Schlick, in his 1932 essay "Positivism and Realism," argued that realism has no place in science because "the 'problem of the reality of the external world' is a meaningless pseudo-problem" (Schlick [1932] 1979, p. 263). Thus, both Carnap and Schlick banished the realism issue from the field of meaningful questions.

However, one must be careful to not overgeneralize this estimation. To be sure, the characterization of the realism issue as a pseudo-problem forms one of the building blocks of the logical empiricist *critique of metaphysics* (see Friedman 2007 and Creath 2014). Yet it must be taken into account that both Carnap and Schlick, while rejecting metaphysical realism, emphatically argued for a non-speculative, *empirical*, realism. More precisely, both Carnap and Schlick thought

of the outer-world hypothesis (the hypothesis of objects existing independently of our consciousness) as meaningless. At the same time, though, they welcomed a realistic interpretation of the empirical statements of science. Thus, Carnap, in his *The Logical Structure of the World* (first published in 1928), points out:

> The realistic language, which the empirical sciences generally use, and the constructional language have actually the same meaning: they are both neutral as far as the decision of the metaphysical problem of reality between realism and idealism is concerned. It must be admitted that, in practice, linguistic realism [*sprachlicher Realismus*], which is very useful in the empirical sciences, is frequently extended to a metaphysical realism; but this is a transgression of the boundaries of science [. . .]. (Carnap [1928b] 1968, pp. 86–87)

In a similar vein, Schlick in "Positivism and Realism" argued that positivism and realism are "not opposed" ([1932] 1979, p. 283) as long as the limits of experience are not transgressed. He even went so far as to contend that anyone who acknowledges the logical empiricist verification principle "must actually be an empirical realist" (ibid.).

Given these qualifications, it remains a large1y open question what exactly was implied by the sort of empirical realism proposed by Carnap and Schlick. Certainly, for Schlick, empirical realism has to be understood in Kantian terms (see Schlick [1932] 1979, Sect. III). However, according to Kant himself, empirical realism is dependent on transcendental *idealism* and thus quite far away from both scientific realism and logical empiricism (see in this regard also Neuber 2014). Be that as it may, the important point to notice is that neither Carnap nor Schlick rejected realism unreservedly. Their rejection was confined to *metaphysical* realism, leaving enough space for accepting the *raison d'être* of a realistic interpretation of the language of science.

3 Realism as a Problem of Language

The work of spelling out this sort of interpretation, though, was left to others. The first one to be mentioned in this connection is Hans Reichenbach who, in his seminal *Experience and Prediction* from 1938, elaborated on the idea that the language of science be interpreted in realistic terms. Reichenbach's frame for designing such a scientific realist account was the theory of meaning, i.e., semantics. What he proposed was a "probability theory of meaning" (see Reichenbach 1938, § 7), which he thought was strong enough to incorporate a semantics for theoretical terms, such as 'atom,' 'electromagnetic field,' etc. (see ibid., § 25)

Reichenbach's conception has been subject of extended investigation by various scholars (see, for example, Salmon 1999a, Putnam 2001, Psillos 2011a, Sober 2011). The crucial point in this conception is the assumption of a *surplus meaning* of theoretical terms. That is, in Reichenbach's view the meaning of theoretical terms is not exhausted by their being reducible to an observational evidence base. Rather, they are invested with an autonomous dimension of explanatory impact, which, Reichenbach maintained, could be elegantly captured by a probabilistic

theory of inductive inference. More precisely, Reichenbach – in the context of his famous 'cubical world' analogy (see Reichenbach 1938, § 14) – pointed out that the existence of theoretical ('unobservable') entities can be inferred inductively by searching for the causes of (regularly occurring) observable effects (like, for example, the tracks in a Wilson cloud chamber). The inferred entities, which Reichenbach called "illata" (see ibid., p. 212), had the status of independently existing things, and their relation to immediately observable entities – Reichenbach called them "concreta" (ibid.) – was that of a "probability connection." Or, as Reichenbach explained by referring to the example of atoms:

> Since all observable qualities of the macroscopic bodies are only averages of qualities of the atoms, there are no strict inferences from the macroscopic bodies to the atom but only probability inferences; we have, therefore, no equivalence between statements about the macroscopic body and statements about the atoms but only a probability connection. (ibid., p. 216)

All of this suggests a strong commitment to the scientific realist agenda. Relying on the specification of the basic scientific realist theses, as it has been provided by Stathis Psillos in his *Scientific Realism: How Science Tracks Truth* (see Psillos 1999, pp. xix–xxi), Reichenbach's position might be summarized as follows: On the ontological level, the independent existence of theoretical entities (such as atoms) is assumed; on the semantic level, we have a theory of meaning for theoretical terms, namely the probability theory of meaning; on the epistemological level, it is assumed that theoretical entities (and their causal properties) are inductively accessible. In short, Reichenbach endorsed all of the central features of modern scientific realism.

However, there are problems lurking in the background. The most obvious of these problems has to do with Reichenbach's interpretation of probability. As is well known, Reichenbach in *Experience and Prediction* defended a *frequency* interpretation of probability (see Reichenbach 1938, §§ 32 and 38). Yet it is by no means clear how by invoking frequencies of observable events ('objective probabilities') the inference to unobservable entities like atoms could be justified. Reichenbach's logical empiricist fellow Herbert Feigl made exactly this point, arguing that

> [t]he crux of the problem lies in the justification of applying the concept of inductive probability to the inference from the directly verifiable to directly unverifiable assertions. Any straightforward frequency interpretation of probability could serve here only if the success frequencies of such inferences were ascertainable. This is outright impossible if independent access to the "Illata" is barred. [...] [T]he legitimacy of applying the probability concept to the whole realistic frame, instead of merely to inferences within it, remains painfully questionable. (Feigl 1950a, p. 53)

The same objection had already been raised by Ernest Nagel (see Nagel 1938, p. 271 and Nagel 1939, p. 237–38). It essentially amounts to the observation that the realistic framework must already be in place in order to make inductive inferences to unobservable entities work. Accordingly, Reichenbach's probability theory of meaning "requires the realist framework and cannot be a proof of it" (Psillos 2011a, p. 37).

This is, however, not the proper forum to examine how Reichenbach's argument for scientific realism could be improved by modifying his account of probability (for an interesting attempt, see Sober 2011). Nor is it my concern to dwell on Reichenbach's later work and on his famous "principle of the common cause" (see in this connection Reichenbach 1956 and the reconstruction in Salmon 2005, pp. 24–25). What should be kept in mind, however, is that the 'probabilistic argument' brought forward in *Experience and Prediction* should be seen as a forerunner of more recent – 'naturalistic' – conceptions of the scientific realistic approach (see in this connection, for example, Boyd 1982, 1983 and Salmon 1984). This is not to say that Reichenbach's probabilistic argument is beyond debate (for a critical discussion, see Psillos 2011a). But it seems to be clear that by focusing on the concept of probability Reichenbach paved the way for mid- and late twentieth-century articulations of the scientific realist agenda.

However, as already indicated, the probabilistic argument cannot be further discussed in this paper. Rather, I wish to take a closer look at Herbert Feigl's approach toward the realism issue, as he outlined it in his essay "Existential Hypotheses: Realistic versus Phenomenalistic Interpretations" (first published in 1950). It is no exaggeration to state that this essay was Feigl's most important contribution to the debate over realism. In its essence, it contained a (language-based) 'pragmatic argument' which must be strictly distinguished from Reichenbach's probabilistic argument.

To begin with, like Reichenbach, Feigl intended to provide us with an affirmative (or constructive) treatment of the realist idea. Furthermore, Feigl, again like Reichenbach, based his argumentation on semantics. By taking semantics seriously, he maintained, "[t]he glib and easy dismissal of the issue as a pseudo-problem will no longer do" (Feigl 1950a, p. 36). Accordingly, what Feigl basically intended was, as he claimed, a "rapprochement" between a "critical phenomenalism (or operationism)," on the one hand, and a "critical (or empirical) scientific realism," on the other (ibid., p. 41). Feigl called the resulting position "Semantic Realism" (ibid., p. 50) and demarcated it from what he called "Probabilistic Realism" (ibid., p. 52). The latter point of view was, as Feigl explicitly remarked, the one defended, among others, by Reichenbach (see ibid., p. 45). As already pointed out, Feigl refused Reichenbach's frequentist interpretation of probability. More generally, he repudiated the entire probabilistic approach. According to Feigl, scientific realism with its "existential hypotheses" concerning theoretical entities could not be justified inductively. Quite the other way round:

> Instead of justifying the surplus meaning of existential hypotheses and hypothetical constructs (Reichenbach's "illata") by means of inductive probability, I suggest that we justify the conceptual frame of the realistic language by its entailed consequence; viz. by showing that only within such a frame it makes sense to assign probabilities to existential hypotheses. (ibid., p. 54)

Thus, in Feigl's view, we first have to establish the realist framework and we then are in a position to raise questions about the probability of specific existential hypotheses concerning theoretical entities. Or, as he argued at another place:

> The customary probabilistic realism in trying to justify "transcendent" hypotheses on the basis of experimental findings has put the cart before the horse. Only after the introduction of the realistic frame can we legitimately argue inductively either from the theory to the outcome of as yet unperformed experiments; or vice versa from the results of experiments to *specific* postulates of the theory. (Feigl 1950b, p. 195)

Thus, according to Feigl, it is "the presupposed introduction of the realistic frame, i.e. the semantic-realistic interpretation of the theory" which "furnishes the very possibility of a theory that is inductively fruitful" (ibid.).

But how, then, can the adoption of the realist framework itself be motivated? As Psillos has correctly observed, answering this question from the perspective of Feigl "is, *ultimately*, a matter of convention" (Psillos 2011b, p. 308). That is, for Feigl, realism is dependent on a foregoing *conventionalist decision*. Consequently, realism cannot be justified naturalistically, but only in a quasi-transcendental manner. By 'quasi-transcendental' I mean the assumption that we cannot directly refer to theoretical (unobservable) entities but that we first have to reflect on the 'conditions of the possibility' of drawing inductive-probabilistic inferences. As in the case of Feigl, this assumption is only *quasi*-transcendental because it is not related to (truth-conducive) statements but only to (action- and decision-relevant) conventions and thus rather 'regulative' than 'constitutive' in the original Kantian sense. As such, it serves as the basis of an essentially *pragmatic* argument. Or, in Feigl's own words:

> The introduction of new *basic* and *irreducible* concepts (as, for example, in electromagnetics during the last century) may be reconstructed as an expansion of the empirical language. Only after our language has thus been enriched, can we significantly assign probabilities (degrees of confirmation) to specific predictive or explanatory hypotheses. The step of expansion of language cannot itself be justified on the grounds of probability, except perhaps in the sophisticated pragmatic sense of the question: Will this expansion be methodologically fruitful? (Feigl 1950a, p. 57)

After all, it is the insight in the "need for definitional or conventional stipulation" (ibid., p. 54) by which, according to Feigl, the realist enterprise is motivated in the first place. The 'condition of the possibility' of the realist program lies outside the reach of the realist program itself. Or as Psillos has aptly put it, for Feigl, there is "no *ultimate* argument for the adoption of the realist framework" (Psillos 2011b, p. 303). Ontic questions are pragmatic "framework-questions" (ibid.), and framework-questions must be decided by convention before any specific existential hypothesis concerning theoretical entities can be evaluated.

The problematic aspects of this quasi-transcendental, convention-based, justification of scientific realism have been discussed elsewhere (see Neuber 2011, 2014). To put it in a nutshell, Feigl's approach seems not to go beyond the later Carnap's ontological 'neutralism' (see Carnap 1956 [1950]). On the other hand, it must be seen that Feigl's contribution formed an autonomous variety of 'realistic claims in logical empiricism.' Especially his contention that theoretical terms have "factual reference" (Feigl 1950a, p. 48) distinguished his 'semantic realism' as a remarkable deviation from early, verificationist, accounts of logical empiricism. However, as Carl Gustav Hempel (1950, pp. 172–73) pointed out in his critique of Feigl's view, the very conception of factual reference fell victim to other restrictions within

the logical empiricist agenda. After all, Feigl's semantic realism boiled down to the charge that theoretical statements be "indirectly confirmable" (Feigl 1950a, p. 57). Their 'factuality' was tied to directly confirmable observation statements, the systematic function of which, Feigl maintained, provided "a maximum of nomological coherence by means of a minimum of hypothetical construction" (ibid.). No doubt that an instrumentalist (or operationist) would have embraced this point of view, all the more since Feigl repeatedly claimed that the realist frame itself was nothing but a "basic convention" (ibid.) that could be "justified only instrumentally" (Feigl 1950b, p. 195). It was for this reason that Hempel did, as he concluded, "not feel convinced that reliance on the problematic concept of the factual referents of theoretical constructs is necessary or even helpful in an attempt to achieve a comprehensive and coherent theoretical account of scientific method and scientific knowledge" (Hempel 1950, p. 173). In fact, also at a larger scale Feigl's plea for scientific realism did not come to fruition. The 'pragmatic argument' he offered could hardly convince the philosophy of science community (for the details of this diagnosis, see Neuber 2011; for a recent vindication of Feigl's pragmatic argument, see Psillos 2011b).

Hempel's critique of Feigl's interpretation of the status of theoretical concepts formed the point of departure for a third variety of 'realistic claims in logical empiricism.' In his guiding paper "The Theoretician's Dilemma: A Study in the Logic of Theory Construction," first published in 1958, Hempel focused on the *purpose* of scientific theory construction. As he saw it, the principle aim of building theories was "systematization." Conceiving of theories as axiomatized systems (see Hempel 1958, p. 46), Hempel confronted the reader with the following – straightforwardly anti-realist – line of reasoning:

> If the terms and principles of a theory serve their purpose they are unnecessary, as just pointed out, and if they don't serve their purpose they are surely unnecessary. But given any theory, its terms and principles either serve their purpose or they don't. Hence, the terms and principles of any theory are unnecessary. (ibid., pp. 49–50)

This argument is called by Hempel the *theoretician's dilemma* (ibid., p. 50). However, it must be said that the dilemma's second horn is trivial, while the first horn of the dilemma is in need of comment. It seems to represent the view of the 'sophisticated' anti-realist. Thus, it could be agreed upon that theoretical concepts and statements serve their purpose if they establish nomological connections among observable phenomena. But then, the sophisticated anti-realist could argue, theoretical concepts and statements can be dispensed with since they are replaceable by concepts and statements that directly refer to the realm of observable phenomena. Logical techniques such as Craig's theorem or the Ramsey-sentence apparently substantiate this abstract claim (see ibid., Sect. 9).

It is not very difficult to see that Feigl's account of scientific realism is suspiciously close to the sophisticated anti-realist's conception. No wonder, then, that Hempel rejected Feigl's point of view as unconvincing. But what was his alternative? As Hempel comprehensively points out in "The Theoretician's Dilemma," one must distinguish between two types of systematization: deductive and inductive

systematization. While in the context of deductive systematization theoretical terms are dispensable, they are *in*dispensable for the purposes of inductive systematization. According to Hempel, theories are *partially interpreted systems*, i.e., systems of concepts and statements that cannot be entirely reduced to the observational evidence base. He therefore is convinced that "the Ramsey-sentence associated with an interpreted theory T' avoids reference to hypothetical entities only in letter – replacing Latin constants by Greek variables – rather than in spirit" (ibid., p. 81). In fact, Hempel maintains, "Ramsey-sentences provide no satisfactory way of avoiding theoretical concepts" (ibid.). This comes as no surprise, since it is theoretical concepts that are needed for the sake of inductive systematization. However, as Hempel argues in direct contradistinction to Feigl, "*semantics* does not enable us to decide whether the theoretical terms in a given system T' do, or do not, have semantical, factual, or ontological reference" (ibid., p. 82: my emphasis). From a purely semantic point of view, the referent of *any* term can be specified, given that our metalanguage is rich enough. Therefore, Hempel concludes, "we have to look elsewhere for criteria of significance for theoretical terms and sentences" (ibid.)

What are these criteria? According to Hempel, we just have to know the *rules* by which sentences of the basic observational vocabulary, V_B, are inferred from sentences containing theoretical terms. This exactly is provided by the procedure of partial interpretation, and we thereby at the same time obtain a workable conception of how inductive relations are established among observable phenomena Thus, given a theoretical hypothesis H_T entails observational consequences OC_1, OC_2, ..., OC_n, we can inductively infer that H_T is true. Further, given that H_T entails a new confirmable prediction OC_{n+1}, we are obviously entitled to conclude that H_T is indispensable because the derivation of OC_{n+1} rests – in an essential way – on the assumption that the inductively obtained hypothesis H_T is true. This is Hempel's way out of the theoretician's dilemma. He claims to have convincingly shown that it starts with a false premise, namely that theoretical terms and sentences, if they serve their purpose, are unnecessary (see ibid., p. 87). Accordingly, for Hempel, a realist interpretation of science is justified. Theoretical systems can, on that basis, be regarded as significant, and the factual reference of theoretical terms can be captured by the following *deflationist account of truth*: "To assert that the terms of a given theory have factual reference, that the entities they purport to refer to actually exist, is tantamount to asserting that what the theory tells us is true; and this in turn is tantamount to asserting the theory." (ibid., p. 84) Thus, when we assert that the elementary particles of contemporary physical theory exist, we assert the truth of the (partially interpreted) physical theory of elementary particles. Moreover, Hempel maintains that on his account the basic tenets of empiricist philosophy can be kept up. In particular, he is eager to tie the theoretical vocabulary to the basic observational vocabulary. The factual reference of theoretical terms is, in Hempel's view, straightforwardly implied by the theory's being true, and the theory's being true can be determined by "an empirical investigation of its V_B-consequences" (ibid., p. 85)

It is hard to see why Hempel's approach should mark a step beyond the point of view defended by Feigl. To be sure, the insight in the indispensability of

theoretical terms is a necessary condition for holding a realist position in the philosophy of science. But the 'indispensability argument' as such is by no means sufficient. Scientific anti-realists could concede the indispensability of theoretical terms but at the same time deny their factual reference. They could, in other words, admit that theoretical terms are necessarily needed for the sake of inductive systematization, but (without becoming bogged down in contradictions) contend that their function is exhausted by this purely systematizing role. That is to say, what is missing in Hempel's approach is an independent argument for the claim that theoretical terms factually refer to (independently existing) *theoretical entities*. Without such an additional argument, Hempel's conception remains open for non-realist reformulations in the spirit of the later Carnap's 'external/internal questions'-point of view (for a fuller discussion of this see Salmon 1999b, pp. 336–37 and Salmon 2005, pp. 26–28).

4 The Invariantist Alternative

By making the factual reference of theoretical terms derivative from theoretical truth, Hempel remains, despite his own contention, within the realm of semantics and thereby within the interpretation of the realism issue as a problem of language. Like Feigl, he finally ends up with a severe empiricist restriction: theoretical truths – and with them theoretical terms – must be essentially tied to the foundation of observational evidence. But why then is this argument operating along realist lines? Would it not suffice to focus on the observational (experimental) adequacy of scientific theories? Or, as Ernest Nagel put it in his critique of Feigl's "Existential Hypotheses:"

> [W]hether one assumes existential hypotheses to be translatable into the language of direct observation, or construes them as elements in a complex symbolic apparatus whose function is to establish systematic relations between experimental data, in either case it seems quite intelligible to assert that a hypothesis is in agreement with a given body of evidence to some specified extent. (Nagel 1950, p. 181)

Why, then, should logical empiricists allow for a realist reading of science at all?

A possible answer to this question is that such a realist reading is *demanded by science itself*. In a certain sense, among the logical empiricists it was Reichenbach who, by invoking the concept of probability, initiated such a non-transcendental, *naturalistic*, approach to science and scientific theory construction. However, the one who articulated this approach most potently was (at least in my opinion) Eino Kaila. According to Kaila, the realism issue is definitely *not* a problem of language. In his view, the problems of philosophy concern, ultimately, scientifically described reality rather than (the quasi-transcendental) questions of 'language engineering.' Thus, as early on as in 1930, in his *Logistic Neopositivism* (a critique particularly of Carnap's *Aufbau*), Kaila declares that "the 'realist language' of science is actually far more than a mere manner of speaking: it is the expression of the living *soul*

of science" (Kaila [1930] 1979, p. 4). To be sure, this could be interpreted as the articulation of (a certain variant of) 'bad metaphysics.' However, what Kaila intends to clarify is that a mere reflection on the language of science is not enough in order to account for the *empirical content* of scientific theories. In other words, according to Kaila it is impossible – or, better, irresponsible – to ignore the 'material mode of speaking' (*inhaltliche Redeweise*) and to restrict philosophical analysis to the 'formal mode of speaking' (*formale Redeweise*). This – essentially Carnapian – strategy would, in Kaila's view, be *not empiricist at all*. It would rather amount to an empirically empty (almost 'scholastic') formalism. On the other hand, Kaila goes not so far as to embrace the idiom of speculative metaphysics. As he stresses in his recently translated *Inhimillinen tieto* (published first in 1939), "the assumption that 'behind' experience there is another, intellectually more perfect world" (Kaila [1939] 2014, p. 29) is empirically not confirmed and hopelessly "imprecise" (ibid.). Thus it is empirical *scientific* realism rather than metaphysical realism by which the Kailaian point of view is driven. Or, as Niiniluoto has once put it:

> Kaila had high respect for the exact philosophical method of the Vienna Circle. Therefore, he strived for a careful formulation of the realism issue, one that would satisfy the critical demands of the new logical empiricism. But it was clear that Kaila – the philosopher of nature who wanted to solve the riddle of reality – could not follow the "linguistic turn" of Analytical Philosophy: for him the deepest problems of philosophy concern *reality* rather than *language*. (Niiniluoto 1992, p. 103)

Kaila, who stood in close contact both to Reichenbach and to the members of the Vienna Circle (see Manninen 2012), grounded his approach to science on two major principles: the principle of *testability* (see esp. Kaila [1936] 1979, pp. 62–63) and the principle of *invariance* (see esp. Kaila [1941] 1979, pp. 149–162). While he characterized the first principle as the "principle of logical empiricism" ([1936] 1979, p. 62), the second principle served, as it were, as his *criterion of reality*. Accordingly, Kaila's point of view should be conceived of as a *fourth* variety of 'realistic claims in logical empiricism.' As such, it can be best characterized as being based on an 'invariantist' – directly science-related (and thus not metaphysically motivated) – argument. Hence it is, I claim, appropriate (and justified) to see in Kaila the most explicit proponent of a realistically inspired variant of logical empiricism.

In order to substantiate this contention, it is advisable to first have a short glimpse at Kaila's principle of testability. As he points out in his monograph *On the Concept of Reality in Physical Science: Second Contribution to Logical Empiricism*, first published in 1941, it is "measurement statements" by which the theoretical hypotheses of physics are empirically tested. Kaila writes:

> [T]he principle of physical testability, which defines empirical statements as 'physical', states that the real content of any physical statement [. . .] consists in the set of measurement statements which are derivable from the statement (in connection with given data). A statement which does not have any such real content is by definition not a physical statement. This principle is implied by the requirement that the singular empirical statements of physics (the basic statements) be exclusively measurement statements. (Kaila [1941] 1979, p. 184)

Had Kaila in earlier writings demanded that theoretical statements be *translatable* into the language of observation (see Niiniluoto 2012, pp. 79–80), he now felt content with their being testable by executing measurements: "[T]he assumption of translatability is not necessary [...]; testability would suffice" (Kaila [1941] 1979, p. 143). Moreover, Kaila clearly saw the need for *idealization* in scientific theory construction and therefore contended that "no theory is decidable, verifiable or falsifiable, in the strict sense; there is decidability only in a certain 'relaxed' or 'weakened' sense; this, however, is testability" (ibid., p. 162).

Kaila's second principle, the principle of invariance, may be portrayed as the very core of his entire philosophical conception (see von Wright 1992, pp. 80–81). In a nutshell, this principle implies that whenever we talk about (both scientific and everyday) reality, we refer to 'invariances.' "There is knowledge only," Kaila maintains, "when some similarity, sameness, uniformity, analogy, in brief, *some 'invariance' is found and given a name*. In knowledge, we are always concerned with 'invariances' alone" (Kaila [1941] 1979, p. 131). As Kaila further points out, the discovery of invariances always goes along with the establishment of a certain structural identity (or isomorphism). In Kaila's own words:

> [I]f one succeeds, e.g., in giving for some domain an account which is in some sense 'unified', then we have the discovery of an 'invariance'; some characteristic or other of higher or lover conceptual level will then have been shown to be invariant with respect to a permutation of the places of the domain. Likewise, e.g., in any formal analogy, structural identity, isomorphism between two different domains, there is also some logically or mathematically definable 'structure', e.g., an equation, that is invariant with respect to the interchange of these domains. (ibid, p. 151)

All of this amounts to a 'structural realist' account of science and scientific theory construction. According to Kaila, it is invariant structures that are captured and described by our best corroborated theories of physical reality. One can even go as far as to say that, for Kaila, physical reality is *nothing but* invariant structures. "*The 'real'*," Kaila declares, "*is what is in some respect (relatively) invariant*" (ibid., p. 185). It is *relatively* invariant because, in Kaila's view, we have, according to the respective *degree* of invariance, different *layers* of reality. Thus Kaila provides us with some sort of ontological hierarchy which extends from perceptual reality to (thing-like) everyday reality and eventually to what is called by him 'physico-scientific reality.' Or, in his own words:

> The physical reality of everyday is a system of invariances of experience, in which a large part of the phenomena is adjudged as 'illusion' and eliminated. Physico-scientific reality is the system of higher invariances of everyday reality, in which again a large part of the latter reality is adjudged as 'illusion' and eliminated. [...] [P]hysico-scientific reality, which is represented by the system of real-descriptions, is in logical respects the highest reality we can attain. (ibid.)

So much should have become clear by now: Kaila's invariantism delivers an independent, non-linguistic, argument for a scientifically realist articulation of the logical empiricist program. According to him, invariance is not inherent in our language but an immanent feature of physical reality (of which our language systems are a part of). Conceived that way, Kaila's invariantist alternative implies that

"physical and scientific objects are objective, independent of us and our perceptions" (Niiniluoto 1992, p. 113). Yet this does not imply that Kaila fell back into the idiom of speculative metaphysics. To be sure, the 'best' invariances we have in science (such as Noether's theorem in classical mathematical physics, comparable invariant theorems in quantum mechanics, etc.) are *formulated in mathematical languages* (governed by certain logical and non-logical rules). Therefore Kaila's realistic, non-linguistic, interpretation of the invariance concept might appear to be too far-reaching in terms of 'ontological commitment' (at least from an empiricist point of view). However, in order to account for the *empirical content* of the invariance concept (and its various applications) the realistic interpretation is the only plausible way to go, since invariances are nothing that can be directly observed. On the other hand, it cannot be denied that by formulating principles of invariance something about empirical reality is intended to be conveyed. In short, Kaila's invariantist argument is sound, provided we accept its status as an answer to the question of how the mathematized language of theoretical physics can be *empirically interpreted*.

However, it still remains to be shown how the principle of testability and the principle of invariance are tied to each other. Concerns of space prevent an extended discussion of this point in this investigation, but suffice it to note that Kaila's *theory of measurement* is destined to achieve the desired solution. According to this (thoroughly anti-conventionalist) theory, it is *metrical relations* that are subject of the application of the principle of testability. Metrical relations, in turn, are the building blocks of Kaila's invariantist ontology. They are what, in the first place, render measurement possible and, thus, are to be seen as "elementary facts which must be present *independently* of measurement" (Kaila [1941] 1979, p. 200; my emphasis). Thus according to Kaila, the (physically) real is what can be measured, and what can be measured are invariant systems of relations, *viz.* structures. Consequently, his invariantist approach comes very close to structural realism. The following passage from his book on human knowledge might serve to corroborate this claim:

> Kant argued that knowledge pertains to appearances only and not to 'things-in-themselves'. And yet he clearly thought that there is an isomorphic relation between appearances and things-in-themselves. That is to say, appearances are representations of things-in-themselves; they share a structure, although, according to Kant, that structure is realized in material that is completely different in the two cases.
>
> We can see therefore that it is wrong to say that we know *nothing* of things-in-themselves: after all, we do know their structure. And if the extreme view turned out to be correct that our knowledge is in the last analysis just a matter of mere representation, we would have to say that we know just as much about things-in-themselves as we do about appearances. (Kaila [1939] 2014, p. 14)

Interestingly enough, in the footnote pertaining to this passage Kaila refers the reader to Bertrand Russell's (structuralist) *Introduction to Mathematical Philosophy* from 1919 (see, in this connection, especially Russell 1919, p. 61). At any rate, for Kaila, the structure of Kantian things-in-themselves is knowable because invariant structures are what can be measured.

5 Concluding Remarks

By way of conclusion, it is instructive to note that the four described varieties of realistic claims in logical empiricism evolved from a rather defensive 'empirical realist' account of science and nature (Carnap and Schlick) to a significantly more offensive articulation of realistic intuitions within the realm of semantics. Reichenbach's probabilistic, Feigl's pragmatic and Hempel's indispensability argument have to be seen against this background. Yet, neither of these arguments is really persuasive. The invariantist argument delivered by Kaila, on the other hand, is more promising. The finally effected fusion of logical empiricist (principle of testability) and structural realist (principle of invariance) components has the potential to stake out a middle path between so-called 'epistemic' structural realism and so-called 'ontic' structural realism (for the details of this distinction, see Ladyman 1998). The principle implications of the resulting variant of a '*metrological*' structural realism have been indicated elsewhere (see Neuber 2012, esp. pp. 374–379); its full systematic exploitation, though, remains the subject of future enquiry. However, it has hopefully become clear in the present paper that Kaila's 'invariantist alternative' must be appreciated as a highly original contribution to the logical empiricist movement and therefore deserves more attention than it has received so far.

Acknowledgments The author would like to thank Joseph J. Kominkiewicz and two anonymous referees who made suggestions that improved this paper significantly.

References

Boyd, R. (1982). Scientific realism and naturalistic epistemology. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1980* (vol. 2, pp. 613–662).

Boyd, R. (1983). The current status of the issue of scientific realism. *Erkenntnis, 19*, 45–90.

Carnap, R. (1956 [1950]). Empiricism, semantics, and ontology. In *Meaning and necessity* (2nd ed., with Supplementary essays pp. 205–221). Chicago: The University of Chicago Press.

Carnap, R. (1968 [1928a]). *The logical structure of the world* (R. A. George, Trans.). London: Routledge.

Carnap, R. (1968 [1928b]). *Pseudoproblems in philosophy* (R. A. George, Trans.). London: Routledge.

Creath, R. (2014). (Anti-)Metaphysics in the thirties: And why should anyone care now? In M. C. Galavotti, E. Nemeth, & F. Stadler (Eds.), *European philosophy of science – philosophy of science in Europe and the Viennese heritage* (pp. 67–76). Dordrecht/Heidelberg/New York/London: Springer.

Feigl, H. (1950a). Existential hypotheses. *Philosophy of Science, 17*, 35–62.

Feigl, H. (1950b). Logical reconstruction, realism and pure semiotic. *Philosophy of Science, 17*, 186–195.

Friedman, M. (1999). *Reconsidering logical positivism*. Cambridge: Cambridge University Press.

Friedman, M. (2007). The *Aufbau* and the rejection of metaphysics. In M. Friedman & R. Creath (Eds.), *The Cambridge companion to carnap* (pp. 129–152). Cambridge: Cambridge University Press.

Hempel, C. G. (1950). A note on semantic realism. *Philosophy of Science, 17*, 169–173.

Hempel, C. G. (1958). The theoretician's dilemma: A study in the logic of theory construction. In H. Feigl et al. (Eds.), *Minnesota studies in the philosophy of science* (Vol. II, pp. 37–98). Minneapolis: University of Minnesota Press.

Kaila, E. (1979 [1930]). Logistic neopositivism: A critical study. In R. S. Cohen (Ed.), *Eino Kaila: Reality and experience. Four philosophical essays* (pp. 1–58). Dordrecht: Reidel.

Kaila, E. (1979 [1936]). On the system of the concepts of reality. A contribution to logical empiricism. In R. S. Cohen (Ed.), *Eino Kaila: Reality and experience. Four philosophical essays* (pp. 59–125). Dordrecht: Reidel.

Kaila, E. (1979 [1941]). On the concept of reality in physical science. Second contribution to logical empiricism. In R. S. Cohen (Ed.), *Eino Kaila: Reality and experience. Four philosophical essays* (pp. 126–258). Dordrecht: Reidel.

Kaila, E. (2014 [1939]). Human knowledge: A classic statement of logical empiricism (A. Korhonen, Trans.). In J. Manninen, I. Niiniluoto, & G. A. Reisch (Eds.), Chicago: Open Court.

Ladyman, J. (1998). What is structural realism? *Studies in the History and Philosophy of Science, 29*, 409–424.

Manninen, J. (2012). Eino Kaila in Carnap's circle. In I. Niiniluoto & S. Pihlström (Eds.), *Reappraisals of Eino Kaila's philosophy* (pp. 9–52). Helsinki: Societas Philosophica Fennica.

Nagel, E. (1938). Review of Reichenbach's *experience and prediction. The Journal of Philosophy, 35*, 270–272.

Nagel, E. (1939). Probability and the theory of knowledge. *Philosophy of Science, 6*, 212–253.

Nagel, E. (1950). A note on semantic realism. *Philosophy of Science, 17*, 169–181.

Neuber, M. (2011). Feigl's 'scientific realism'. *Philosophy of Science, 78*, 165–183.

Neuber, M. (2012). Invariance, structure, measurement – Eino Kaila and the history of logical empiricism. *Theoria, 78*, 358–383.

Neuber, M. (2014). Is logical empiricism compatible with scientific realism? In M. C. Galavotti, E. Nemeth, & F. Stadler (Eds.), *European philosophy of science – philosophy of science in Europe and the Viennese heritage* (pp. 247–262). Dordrecht/Heidelberg/New York/London: Springer.

Niiniluoto, I. (1992). Eino Kaila and scientific realism. In I. Niiniluoto, M. Sintonen, & G. H. von Wright (Eds.), *Eino Kaila and logical empiricism* (pp. 102–116). Helsinki: Societas Philosophica Fennica.

Niiniluoto, I. (1996). *Critical scientific realism*. Oxford: Oxford University Press.

Niiniluoto, I. (2012). Eino Kaila's critique of metaphysics. In I. Niiniluoto & S. Pihlström (Eds.), *Reappraisals of Eino Kaila's philosophy* (pp. 71–89). Helsinki: Societas Philosophica Fennica.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. London/New York: Routledge.

Psillos, S. (2011a). On Reichenbach's argument for scientific realism. *Synthese, 181*, 23–40.

Psillos, S. (2011b). Choosing the realist framework. *Synthese, 180*, 301–316.

Putnam, H. (2001). Hans Reichenbach: Realist and Verificationist. In J. Floyd & S. Shieh (Eds.), *Future past: The analytic tradition in twentieth-century philosophy* (pp. 277–287). Oxford: Oxford University Press.

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge*. Chicago: The University of Chicago Press.

Reichenbach, H. (1956). *The direction of time* (M. Reichenbach Ed.). Berkeley/Los Angeles: University of California Press.

Russell, B. (1919). *Introduction to mathematical philosophy*. London: George Allen and Unwin.

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

Salmon, W. (1999a). Ornithology in a cubical world: Reichenbach on scientific realism. In D. Greenberger, R. L. Wolfgang, & A. Zeilinger (Eds.), *Epistemological and experimental perspectives on quantum physics* (pp. 305–315). Dordrecht: Kluwer.

Salmon, W. (1999b). The spirit of logical empiricism: Carl G. Hempel's role in twentieth-century philosophy of science. *Philosophy of Science, 66*, 333–350.

Salmon, W. (2005). Scientific realism in the empiricist tradition. In P. Dowe, & M. H. Salmon (Eds.), *Reality and rationality* (pp. 19–30). Oxford: Oxford University Press.

Schlick, M. (1979 [1932]). Positivism and realism. In H. L. Mulder, & B. van de Velde-Schlick (Eds.), *Philosophical papers* (vol. 2, pp. 259–284). Dordrecht: Reidel.

Sober, E. (2011). Reichenbach's cubical universe and the problem of the external world. *Synthese, 181*, 3–21.

Stadler, F. (2001). *The Vienna circle: Studies in the origins, development, and influences of logical empiricism*. Vienna: Springer.

Uebel, T. (2007). *Empiricism at the crossroads: The Vienna circle's protocol-sentence debate*. Chicago/LaSalle: Open Court.

Von Wright, G. H. (1992). Eino Kaila's monism. In I. Niiniluoto, M. Sintonen, & G. H. von Wright (Eds.), *Eino Kaila and logical empiricism* (pp. 71–91). Helsinki: Societas Philosophica Fennica.

Patchworks of Laws and Partial Structures

Holger Andreas

1 Patchwork of Laws

In her seminal *How the Laws of Physics Lie* (Cartwright 1983), Nancy Cartwright has levelled an influential attack at the view that nature is governed by universally true, factual laws. *The Dappled World: A Study of the Boundaries of Science* (Cartwright 1999), then, develops an alternative view which is guided by the metaphor of a *patchwork of laws*. This metaphor is well worked out by case studies and general considerations about models, nomological machines, capacities, bridge principles, etc. A particular strength of her investigation lies in that it is not only concerned with abstract, purified laws but with the application of such laws. A logically minded reader, however, finds it desirable to have a more formal elaboration of the non-universalist view of nature. In particular she will wonder whether it is feasible to devise a logical system that is expressive enough to capture scientific reasoning with patchworks of laws. The present paper aims to answer this question in the affirmative and, consequently, expounds the basic elements of such a system.

Before I outline a formal account of patchworks of laws, let me briefly explain why I think the non-universalist picture of science deserves further investigation. Nancy Cartwright expresses her criticism of the universalist picture with different degrees of strength. In her *How the Laws of Physics Lie*, there are bold statements claiming this picture to be simply wrong (Cartwright 1983, p. 46):

> Most scientific explanations use *ceteris paribus* laws. These laws, read literally as descriptive statements, are false, not only false but deemed false in the context of use.

H. Andreas (✉)
University of British Columbia, Okanagan Campus, Kelowna, BC, Canada
e-mail: holger.andreas@lrz.uni-muenchen.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_4

In support of this and related claims, she discusses the problem of composing laws. This problem arises for physical systems that require the application of more than one quantitative law. For example, where a body is subject to both gravitational and electromagnetic forces, the law of vector addition lets us calculate the net effect of component forces each of which is captured by a precise quantitative law. However, Cartwright (1983, p. 63–69) argues, by way of examples taken from fluid dynamics and atomic physics, that laws of composition are often not available.

The Dappled World (Cartwright 1999) criticises the universalist picture in a more sophisticated and less definite manner. At its core, this book advances the following pluralist doctrine (Cartwright 1999, p. 31):

> Metaphysical nomological pluralism is the doctrine that nature is governed in different domains by different systems of laws not necessarily related to each other in a systematic or uniform way; by a patchwork of laws.

How does she argue for this doctrine? As a matter of fact, scientists are actually working with different models, different entities and concepts, depending on the domain of inquiry. There is no single scientific discipline for which a universal base of a few principles and laws could yet be established in such a manner that all more specific laws – also called *phenomenological* by Cartwright – have become derivable from such a base. Cartwright thinks there is a systematic reason for this diversity: all laws, phenomenological and fundamental ones alike, have specific limits of applicability and validity. This is why she holds the doctrine that all laws are *ceteris paribus*, i.e., holding only under conditions that are not made explicit by the law's formulation itself.[1]

Observing a large diversity of domain-specific models, laws, and entities that are not related in an uniform way is a 'far cry' from refuting the universalist picture of science, as Cartwright (1983, p. 10–12) is aware. In response to this objection, she points out that she is striving for a more empiricist account of science that neither idealises away the boundaries of presently available scientific theories nor underestimates issues of applying scientific laws to the world. This, I think, is a very strong point: the universalist picture describes the use and the application of scientific laws as well as up-to-date scientific theories not fully accurate. And it is doubtful whether it will ever do so. "Reality may well be just a patchwork of laws" (Cartwright 1999, p. 34).

There is of course also the issue of normativity. Though the universalist view of nature and science does not describe scientific theories in their present state, we may well understand this view as a normative ideal to be pursued by scientists. Cartwright thinks that this normative understanding of the universalist view of science has harmful implications for science policy. However, the normative dimension of the universalist view is not a concern of the present investigation. I am rather interested in the logical analysis of the boundaries of science as they continue to exist in up-to-date scientific theories.

[1] For a fine-grained account and a classification of *ceteris paribus* laws, see Schurz (2002).

The idea of a patchwork of laws interrelates with two prominent areas in philosophical logic: nonmonotonic and paraconsistent reasoning. As logics of nonmonotonic and paraconsistent reasoning have emerged, non-universal axioms have been recognized as an important means of human reasoning. This suggests using systems of nonmonotonic and paraconsistent reasoning for a logical analysis of the *boundaries of science*. So far, such systems have rarely been exploited for an analysis of scientific reasoning.

Besides reference to paraconsistent and nonmonotonic logics, the present investigation makes use of some elements of the Sneed formalism also known as the structuralist approach to science (Balzer et al. 1987; Sneed 1979). In particular the structuralist's notion of an *intended application* of an axiom, or theory-element, will prove useful in an analysis of scientific reasoning with patchworks of laws. More precisely, I will expound in Sect. 2 a system of paraconsistent reasoning that has emerged from a synthesis of the framework of partial structures by da Costa and French (2003) with a few elements of the Sneed formalism. This synthesis results into a *modular semantics* for axiomatic theories. In Sect. 3, I shall then show that a modular semantics with its network of partial structures is apt to underpin various claims being central to Cartwright's philosophy of science.

2 Modular Semantics for Axiomatic Theories

2.1 Networks of Partial Structures

A modular semantics is obtained by two successive operations upon the standard semantics: (i) the descriptive vocabulary V of the axiomatic theory T is divided into subvocabularies according to the axioms of T, and (ii) these subvocabularies in turn are interpreted by partial structures that represent applications of the corresponding axiom. Let $\alpha_1, \ldots, \alpha_n$ be the axioms of T with corresponding subvocabularies $V(\alpha_1), \ldots, V(\alpha_n)$. The partial structure $\mathcal{A}_{i,j}$ represents the application j of the axiom α_i. For consideration of inter-theoretical relations, T may encompass axioms coming from a number of different scientific theories – so that T would stand for what structuralists call a *holon* of scientific theories. This being said, we can graphically illustrate the basic idea of a modular semantics for axiomatic theories as follows (Fig. 1).

This modularisation of the interpretation of an axiomatic theory will serve as a semantic foundation for using the instances of universal axioms selectively. In brief, this strategy goes as follows: the notion of an interpretation of $\mathcal{L}(V)$ is defined as the composition of "local interpretations" that result from applying axioms of T to certain systems under consideration. (These applications function as modular units.) By default, all applications of all axioms contribute to the interpretation of $\mathcal{L}(V)$. However, in the case of an (internal or external) inconsistency of an application,

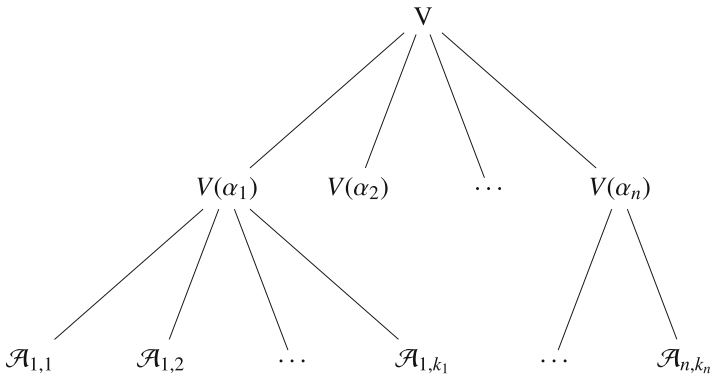


Fig. 1 Modular semantics

this application is ignored for the composition of an interpretation of $\mathcal{L}(V)$. This amounts to selectively accepting the instances of universal axioms of T.

Representing any application of any axiom of T by a separate set-theoretic structure is inspired by the structuralist notion of an intended application as expounded in Sneed (1979) and Balzer et al. (1987). Setting aside the strong semantic orientation of the structuralist approach, we can say that any (empirical or abstract) system for which an axiom α is supposed to hold in virtue of the understanding of α is an intended application of α. For example, a two-body system consisting of the earth and a falling body near the earth's surface qualifies as an intended application of Newton's law of gravitation. x being an intended application of the axiom α does not imply that α actually applies to x. We shall assume that any instance of an axiom uniquely corresponds to an intended application of that axiom, where each such application allows for a set-theoretic representation by a partial structure. Intended applications are the semantic counterparts of instances of universal axioms.

Let us briefly recall the notion of a partial structure as expounded in da Costa and French (2003). A *simple pragmatic structure* is a set-theoretic structure of the form

$$\mathcal{A} = \langle A, R_k, P \rangle_{k \in K}$$

where A is the domain of interpretation, R_k are partial relations, and K is an index set. P is a set of sentences representing putative knowledge. $\mathcal{A}$ thus encodes a partial interpretation of some language L. For an n-ary relation R_k, being partial means that the extension of the relation concept R_k is a superset of the relation R_k.[2]

[2]Partiality of a relation R_k is more precisely accounted for by distinguishing between the positive extension R_k^+, the negative extension R_k^-, and the neutral extension R_k^0. For simplicity, we assume that, for any partial structure $\mathcal{A}$, $R_k^+ = R_k$ and $R_k^- = \emptyset$. That is, for any partial structure $\mathcal{A}$ and any n-ary relation R_k, if an n-tuple x is not a member of $R_k = R_k^+$, it is not a member of R_k^- either.

Why do the partial structures $\mathcal{A}_{i,j}$ form a network in the present modular semantics? Two partial structures may overlap in terms of their domain and of the descriptive vocabulary of which they encode a partial interpretation. That is, (i) the domains A and A' of two corresponding partial structures $\mathcal{A}$ and $\mathcal{A}'$ may be such that $A \cap A' \neq \emptyset$, and (ii) the vocabularies V and V' partially interpreted by $\mathcal{A}$ and $\mathcal{A}'$, respectively, may have the property that $V \cap V' \neq \emptyset$. Partial structures $\mathcal{A}_{i,j}$ are thus interrelated. We may represent these relations by an undirected graph having the following properties: (i) the set of nodes is given by the set of partial structures, and (ii) there is an edge between two nodes iff the corresponding partial structures overlap in terms of their domain and in terms of their vocabulary. Such a graph provides an intuitive means of conceiving a network of partial structures.

2.2 Local Worlds

Let us now go further into the details of a modular semantics for axiomatic theories. Each axiom has a set of applications to empirical or abstract systems of entities. Each application of an axiom is represented by a simple pragmatic structure $\mathcal{A}_{i,j}$, where i indicates the axiom α_i and j the particular application of that axiom:

$$\mathcal{A}_{i,j} = \langle A_j, R_1, \ldots, R_k, \{\alpha_i\}\rangle$$

We consider now the set $W_{i,j}$ of those (model-theoretic) extensions of the partial structure $\langle A_j, R_1, \ldots, R_k\rangle$ which satisfy α_i:

$$W_{i,j} = \begin{cases} Mod(\alpha_i) \cap Ext(\mathcal{A}_{i,j}) \text{ if } Mod(\alpha_i) \cap Ext(\mathcal{A}_{i,j}) \neq \emptyset \\ Ext(\mathcal{A}_{i,j}) \text{ otherwise.} \end{cases} \quad (1)$$

where $Mod(\alpha_i)$ designates the set of models of α_i and $Ext(\mathcal{A}_{i,j})$ the set of model-theoretic extensions of the partial structure given by simple pragmatic structure $\mathcal{A}_{i,j}$.

We have thus obtained a set $W_{i,j}$ of "local worlds" for any application j of any axiom α_i. These local worlds represent admissible extensions of the respective simple pragmatic structure. Obviously, the definition of the set $W_{i,j}$ spells out the semantic consequences of applying an axiom α_i to an empirical or abstract system represented by a simple pragmatic structure $\mathcal{A}_{i,j}$.

This semantic consideration of modular units does, of course, not suffice for an account of scientific reasoning because intended applications are not isolated monadic items. Quite to the contrary, such applications are highly interrelated. Therefrom, the challenge arises of merging the semantic consequences of such

On this assumption, there is no need to notationally distinguish between the positive, the negative, and the neutral extension of a relation R_k in a partial structure.

individual applications of single axioms. For this challenge to meet, we will merge local worlds – which are given by sets $W_{i,j}$ defined by (1) – to build global worlds, i.e., interpretations of the global language $\mathcal{L}(V)$.

2.3 Global Worlds

To study ways of constructing global worlds out of local ones, let us consider the set $\mathbf{W}$ of sets $W_{i,j}$ of local worlds:

$$\mathbf{W} = \{W_{i,j} \mid \alpha_i \text{ is an axiom of } T \text{ and } j \in J_i\} \tag{2}$$

where J_i is the index set for the applications of the axiom α_i.

Two worlds from different sets $W_{i,j}$ may overlap in terms of their domain and in terms of the vocabulary whose interpretation they encode, as was observed above. Two such worlds, thus, may or may not be consistent with each other in the sense of being interpretations that do agree or do not agree wherever there is an overlap between them. The challenge arising here is to construct a global world out of mutually consistent local ones. For this to be achieved, we need to construct choice sets W_c of $\mathbf{W}$ such that for any $w, w' \in W_c$, w is consistent with w'. Once such a consistent choice set W_c is constructed, it is easy to define therefrom a global world by a componentwise union of the members of W_c. Let us describe this construction of global worlds more precisely. As a first step, we define the notion of consistency between two worlds:

Definition 1 (Consistency between two worlds). Let $w_i = \langle A_i, R_k\rangle_{k\in K_i}$ and $w_j = \langle A_j, R_k\rangle_{k\in K_j}$ be two total structures, or worlds. w_i is consistent with w_j iff, for all $k \in K_i \cap K_j$ and all m-tuples x we have (i) if $x \in (R_k)_{w_i}$ and $x \in A_j^m$, then $x \in (R_k)_{w_j}$, and (ii) if $x \in (R_k)_{w_j}$ and $x \in A_i^m$, then $x \in (R_k)_{w_i}$. $(R_k)_w$ denotes the relation R_k of the structure w and A^m the m-ary Cartesian product of A.

Definition 2. $\uplus$ Let w_i and w_j be two local worlds that are consistent with one another in the sense of Definition 1. $w_i \uplus w_j = w_0$ iff

(1) $w_0 = \langle A_0, R_k\rangle_{k\in K_0}$,
(2) $A_0 = A_i \cup A_j$,
(3) $K_0 = K_i \cup K_j$,
(4) for all $k \in K_i \cap K_j$, $(R_k)_{w_0} = (R_k)_{w_i} \cup (R_k)_{w_j}$,
(5) for all $k \in K_i \setminus K_j$, $(R_k)_{w_0} = (R_k)_{w_i}$,
(6) for all $k \in K_j \setminus K_i$, $(R_k)_{w_0} = (R_k)_{w_j}$.

This being so defined, we can say that a global world w is obtained by the composition of local worlds of a consistent choice set W_c iff

$$w = \biguplus_{w_i \in W_c} w_i. \tag{3}$$

This construction yields sensible results in cases where an axiom α of T has an internally inconsistent application, i.e., where there is no extension of $\mathcal{A}_{i,j}$ that satisfies α_i. However, if an application of an axiom α is inconsistent with some application of another axiom β (or with another application of α), there is no choice set W_c of $\mathbf{W}$ such that for any $w, w' \in W_c$, w is consistent with w'.

We, therefore, refine the construction of global worlds in Eq. (3) as follows: in place of consistent choice sets W_c of $\mathbf{W}$, we take consistent choice sets of maximally consistent subsets $\mathbf{W}' \subseteq \mathbf{W}$. A set $\mathbf{W}'$ is called *consistent* iff it has a choice set whose members are consistent with one another. From such a set $\mathbf{W}' \subseteq \mathbf{W}$ we obtain admissible substructures of the global language $\mathcal{L}(V)$:

Definition 3 (Admissible substructure of $\mathcal{L}(V)$). Let the set $\mathbf{W}$ of local worlds be defined by Eq. (2). A structure $\mathcal{A}$ is an admissible substructure of the global language $\mathcal{L}(V)$ iff there is a set W_c such that

(1) W_c is choice set of a set $\mathbf{W}' \subseteq \mathbf{W}$,
(2) for all $w, w' \in W_c$, w is semantically consistent with w',
(3) there are no sets W_c' and $\mathbf{W}''$ such that (i) $\mathbf{W}' \subset \mathbf{W}'' \subseteq \mathbf{W}$, (ii) W_c' is a choice set of $\mathbf{W}''$, and (iii) for all $w, w' \in W_c'$, w is semantically consistent with w',
(4) $\mathcal{A} = \biguplus_{w \in W_c} w$.

The notion of a substructure is understood here as follows: w_1 is a substructure of w iff there is a structure w_2 such that $w = w_1 \uplus w_2$.

Admissible substructures of $\mathcal{L}(V)$ need to be extended in a manner that is consistent with the facts represented by all partial structures $\mathcal{A}_{i,j}$, independently of whether some $x \in W_{i,j}$ is a member of a maximally consistent choice set W_c. This requirement is taken into account by defining the notion of an *admissible total structure* of $\mathcal{L}(V)$:

Definition 4 (Admissible total structure of $\mathcal{L}(V)$). w_0 is an admissible total structure of $\mathcal{L}(V)$ iff there are structures w_1 and w_2 such that

(1) $w_0 = w_1 \uplus w_2$,
(2) w_1 is an admissible substructure of $\mathcal{L}(V)$ in the sense of Definition 3,
(3) for all partial structures $\mathcal{A}_{i,j}$, there is a structure $w \in Ext(\mathcal{A}_{i,j})$ such that w is a substructure of w_0.

Having thus defined the global worlds of T, we introduce a specific notion of truth in a modular semantics:

Definition 5 (Modular theoretical truth). Let $\Box$ be introduced by an S5 modal system for the language $\mathcal{L}(V)$, where the set W of worlds is given by the set of admissible total structures of $\mathcal{L}(V)$ as defined by Definition 4. A sentence ϕ of $\mathcal{L}(V)$ is (modularly) theoretically true iff $\Box\phi$.

In less formal terms, a sentence ϕ is (modularly) theoretically true iff it is true in all global worlds that can be built up by maximally consistent compositions of local worlds, where each set of local worlds represents the semantic consequences of

applying a particular axiom to some phenomenon. Based on such an understanding of modular theoretical truth, an inference relation can be defined for networks of partial structure in a relatively straightforward manner:

Definition 6. Let **N** be a network of partial structures and ϕ a formula of $\mathcal{L}(V)$. The notion of an admissible total structure of $\mathcal{L}(V)$ is understood in the sense of Definition 4. ϕ is inferable from **N** – in symbols: $\mathbf{N} \mid\sim \phi$ – iff ϕ is verified by all admissible total structures of $\mathcal{L}(V)$.

2.4 Inconsistency Management

The intuition driving this modular semantics can perhaps most clearly be perceived by an analysis of paradoxical sentences. Suppose l is a liar sentence, which is self-referential and self-contradictory. When applying the T-scheme

$$T(\alpha) \leftrightarrow \alpha \tag{TS}$$

to l, we realise that this application has no theoretical extension that satisfies this scheme. In other words, $Mod(P) \cap Ext(\mathcal{A}) = \emptyset$, where $P = \{(TS)\}$. Hence, by Eq. (1), local structures representing l as true are as admissible as local structures representing l as false. Hence, we will have global worlds that verify l and other global worlds that falsify it. Hence, l is neither true nor false on the semantics of modular theoretical truth. Liar sentences are "gappy".

Similar considerations apply to axiomatic theories in the natural sciences that fail to be applicable to a particular empirical system. Consider the famous application of Newtonian mechanics to the precession of the perihelion of Mercury. The problem of this application is that we are unable to determine the forces acting upon Mercury – by Newton's law of gravitation for the putative distribution of masses in our solar system – such that Mercury's trajectory accords with Newton's second law of motion. Hence, there is no theoretical description of Mercury's motion in terms of Newtonian gravitational forces that accords exactly with Newton's laws of motions. In our semantics, this has the consequence that some global worlds satisfy Newton's equations for the trajectory of Mercury without satisfying Newton's law of gravitation, while others satisfy Newton's law of gravitation for the gravitational forces acting upon Mercury without satisfying Newton's equations. The precession of Mercury's perihelion, therefore, does not contribute to the interpretation of the force function (for the argument Mercury) in the language of Newtonian mechanics. As a consequence of this, certain sentences about Newtonian forces acting upon Mercury are gappy in our semantics in a manner that is comparable to the "gappiness" of liar sentences. For both types of sentences, gappiness arises because there is a global world that verifies the sentence under consideration and another global world that falsifies it.

3 Patchwork of Partial Structures

A modular semantics with its network of partial structure provides a logical analysis of scientific reasoning with patchworks of laws. This is the key claim of the present paper. Although a modular semantics does not capture every aspect of Cartwright's non-universalist view of science, it is expressive enough to furnish this view with a formal semantics. For this to be seen, let us analyse the following set of related claims and conjectures made by Cartwright (1999):

(1) "The abstract theoretical concepts of high physics describe the world only via the models that interpret these concepts more concretely" (Cartwright 1999, p. 4).
(2) "Laws can be true, but not universal. We need not assume that they are at work everywhere, underlying and determining what is going on. If they apply only in very special circumstances, then perhaps they are true just where we see them operating so successfully – in the artificial environments of our laboratories, our high-tech firms, or our hospitals" (Cartwright 1999, p. 37).
(3) For a *ceteris paribus* law, the conditions of validity cannot be cashed out in the language of the scientific theory to which this law belongs (Cartwright 1999, p. 10).

Claim (1) implies that theoretical concepts do not have referents by themselves. Rather, their ability to describe something in the world is bound to applying models to whatever phenomena. Such models are characterized as *interpretative*. Interpretative models mediate between the theory and the world. I suggest that Cartwright's interpretative models may be described as concepts subsuming similar applications of one and the same law or of a set of laws. This understanding squares well with examples given by Cartwright for interpretative models: pendulums, springs, planetary systems in classical mechanics; the central potential, scattering, the Coulomb interaction, and the harmonic oscillator in quantum mechanics.

In our modular semantics, any phenomenon being subject to the application of a law is formally represented by a partial relational structure. This structure specifies, at least partially, the extension of those concepts in terms of which we describe the respective phenomenon but leaves the extension of the theoretical concepts unspecified. The latter concepts become determined only after some theoretical law, or some set of theoretical laws, is applied to the respective phenomenon. Hence, in our semantics, theoretical concepts do not describe anything prior the application of theoretical axioms. The types of applications may well be described by interpretative models, such as a pendulum, a spring, a planetary system, etc., though this description is not part of the formal semantics.[3]

[3]There is no room here for elaborating the distinction between phenomenal and theoretical concepts. I suggest understanding this distinction in a strictly relativised way. See Andreas (2013) for details.

Note that our semantics accounts for the interpretation of the theoretical concepts through the application of laws in a precise and patchwork like manner. For this interpretation consists, essentially, in (i) specifying the local worlds of a simple pragmatic structure (Eq. (1)), and (ii) composing global worlds out of the various sets of local worlds (Definitions 3 and 4). This composition may metaphorically be described as fitting the single patches together into a coherent patchwork. Those patches that cannot be made to fit, must be ignored and cannot be used for the patchwork; they fail to interpret the theoretical concepts. Each patch stands for the application of a law to some phenomenon. Each such patch forms a modular unit in our semantics, and the patchwork qualifies as a modular system.

(2) seems to qualify as a mere conjecture about the scope of laws in science. Alternatively, we may take (2) as part of a proposal for a non-universalist metatheory that is more empiricist, more subtle, and that allows for a more fine-grained analysis of scientific reasoning than the universalist view. On this understanding, the proposal is to say that laws are (i) true where we see them operating successfully, (ii) false where they are unsuccessful, and (iii) devoid of meaning and truth-value in domains where their application is unclear. This distinction is represented by our modular semantics, with the qualification that we do not consider computational limits.[4] In particular, a law has no truth-value at all, according to our semantics, in domains where it is not applied. The distinction between successful and unsuccessful applications of scientific laws can be made fairly precise: an application of a law α_i represented by a simple pragmatic structure $\mathcal{A}_{i,j}$ is successful iff, for all global worlds w (which are defined by Definition 4), there is a structure $w' \in W_{i,j}$ such that (i) w' is a substructure of the global structure w, and (ii) $w' \in Mod(\alpha_i)$. Otherwise, this application of α_i is unsuccessful.

Claim (3) poses a well known problem for standard representations of axiomatic scientific theories in the language of classical logic. It is related to the more general question of whether it is, in principle, possible to explicitly specify the conditions of validity of *ceteris paribus* laws. Even so, Cartwright argues, the language of a given scientific theory does not have the resources to spell out these conditions of validity. Hence, we can conclude, classical logic is unable to explain our use of universal axioms in a manner that retains the standard universal notations of such axioms. For standard notations of scientific laws make no reference to any boundaries of applicability, nor do they indicate how a scientific law is applied.

The present modular semantics retains the general format of scientific axioms, without assuming that these axioms are universally valid. For the axioms $\alpha_1, \ldots, \alpha_n$ making up the axiomatic theory T are assumed to have the surface syntax of universal statements. By considering the applications of each axiom individually and by building up semantic interpretations of $\mathcal{L}(V)$ out of such applications, we

[4] That is, in our semantics, a law is true in a subdomain if it is consistently applied there, even if the satisfaction of the consistency requirement does escape our recognition, so that we may fail to see the law operating successfully.

have been able to use these individual applications *selectively*.[5] Once an application $\mathcal{A}_{i,j}$ is found not to be consistent with some other piece of our beliefs (which is represented by some other simple pragmatic structure), it will not contribute to the interpretation of the global language $\mathcal{L}(V)$. $\mathcal{A}_{i,j}$ is then ignored, which notably does not force us to retract other applications of α_i.

Moreover, there is no danger in our semantics to envision ill-defined applications of an axiom because any application must be formally specified by a simple pragmatic structure. Such a structure represents the phenomenon to which an axiom, or a set of axiom, is applied for the purpose of understanding, explanation, and prediction. This tight connection between an axiom and its applications ensures that the theoretical concepts are related to the phenomena under consideration in a precise way.

If we say that fluid dynamics describes what happens to a thousand dollar bill swept away by the wind (cf. Cartwright 1999, p. 27), no relation between the observable motion of the bill and the theoretical concepts of fluid dynamics is established. It is a vacuous application not only because no predictions are made, but also because the theoretical concepts remain wholly underdetermined. This contrasts, for example, with applications of fluid dynamics in meteorology using the *shallow water model*. Among other things, this model allows us to explain the development of cyclones and anticyclones by way of so-called *Rossby waves*. Such waves figure as theoretical entities; their parameters, such as wave length and frequency, become actually determined in specific applications of the shallow water model (cf. Holton and Hakim 2013, Ch. 5.7).

In sum, our modular semantics allows us to retain the standard universal notation of scientific laws, without implying that this notation states their universal validity. If there happen to be well-defined but invalid applications besides a range of valid applications, our semantics is able to demarcate between the two types. The inference relation of this semantics is such that from the set of axioms of T we can infer any instance of an axiom that represents a valid application, whereas the instances of inconsistent applications are not inferable. Hence, it is logically not a problem any more that the conditions of validity of a scientific law are not specified by the law itself.[6] Our modular semantics for axiomatic theories, therefore, coheres much better with the patchwork view of laws than standard representations of scientific axioms using classical logic.

[5]The selective use of universal axioms is a key idea in various formalisms of nonmonotonic reasoning. The present modular semantics draws on Brewka (1991).

[6]This also achieved by other systems of nonmonotonic reasoning. So far, however, little research has been done on exploiting such systems for an analysis of scientific reasoning.

References

Andreas, H. (2013). Theoretical terms in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2013 ed.). http://plato.stanford.edu/entries/theoretical-terms-science/

Balzer, W., Moulines, C. U., & Sneed, J. (1987). *An architectonic for science. The structuralist program*. Dordrecht: D. Reidel.

Brewka, G. (1991). Belief revision in a framework for default reasoning. In *Proceedings of the Workshop on The Logic of Theory Change* (pp. 602–622). London: Springer.

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

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

da Costa, N., & French, S. (2003). *Science and partial truth*. Oxford: Oxford University Press.

Holton, J. R., & Hakim, G. J. (2013). *An introduction to dynamic meteorology*. Oxford: Academic.

Schurz, G. (2002). Ceteris paribus laws: Classification and deconstruction. *Erkenntnis, 57*(3), 351–372.

Sneed, J. (1979). *The logical structure of mathematical physics* (2nd ed.). Dordrecht: D. Reidel.

Part II
Social Epistemology, Rational Choice Theory and Public Policy

Social Epistemology, Debate Dynamics, and Truth Approximation

Gustavo Cevolani

1 Introduction

Social epistemology tackles the classical problems of knowledge, truth and rationality by exploring those social processes, methods and practices—like testimony, authority, belief aggregation or the search for consensus—by which "epistemic subjects" interact with other agents who influence, and are influenced by, their beliefs (Goldman 2010). In particular, "veristic" social epistemology (Goldman 1999; Kitcher 1993)—as opposed to constructivist approaches in the sociology of knowledge—aims at analyzing and evaluating such practices with respect to their tendency to disseminate true beliefs in communities of truth-seeking agents (both scientists and laymen). Within the veristic approach, the *truth conduciveness* of social practices—i.e., their effectiveness in promoting epistemic progress, construed as increasing approximation to the truth, within the relevant community—is thus the crucial issue.

In this paper, I address the issue of truth conduciveness by focusing on the "theory of dialectical structures" recently developed by Gregor Betz (2013). This theory aims at investigating the role of controversial argumentation in consensus formation and truth approximation. More specifically, the theory studies, via computer simulations, how the opinions of a group of agents change as arguments are introduced in the debate, and whether the group reaches a consensus or makes progress toward the truth as a result. As Betz (2013, pp. 39–40) notes, his approach is closely related to the philosophical accounts of verisimilitude or truthlikeness, and in particular to the "basic feature" approach of Cevolani et al. (2011). In

G. Cevolani
Department of Philosophy and Education, University of Turin, via S. Ottavio 20, 10124 Turin, Italy
e-mail: g.cevolani@gmail.com

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_5

Sects. 2 and 3, I briefly present the basic feature approach and the theory of dialectical structures, respectively; Sects. 4 and 5 focus on the formal and conceptual relationships between these two accounts; finally, in Sect. 6, some general remarks on the issue of truth conduciveness in veristic social epistemology are offered.

2 Truth Approximation as the Goal of Inquiry

Intuitively, theory A is verisimilar (or truthlike) if it is close or similar to the whole truth about a given target domain. Thus, A may be false but still a good approximation to the truth, and even a better approximation than another (true or false) theory. This notion was originally introduced by Karl Popper (1963, ch. 10) in order to defend the idea that progress can be explained in terms of the increasing verisimilitude of scientific theories. According to Popper, such theory-changes as that from Newton's to Einstein's theory are progressive because, although the new theory is, strictly speaking, presumably false, we have good reasons to believe that it is closer to the truth than the superseded one: increasing verisimilitude is the key ingredient for progress.[1]

The main idea underlying the basic feature approach (Cevolani et al. 2011, 2013) is that the verisimilitude of a theory can be defined in terms of the balance of the true and false information that it conveys about the "basic features" of the target domain. Suppose that such domain is described by a finite propositional language $\mathscr{L}_n$ with n logically independent atomic sentences $a_1, a_2, \ldots, a_n$. Then, its basic features are described by the so called basic sentences or literals of $\mathscr{L}_n$, i.e., by the atomic sentences and their negations. A non-contradictory conjunction of m basic sentences, with $m \leq n$, will be called a "conjunctive theory" ("c-theory", for short). The 3^n c-theories of $\mathscr{L}_n$ include (for the special case $m = n$) the so called constituents, which are consistent conjunctions of n basic sentences. Constituents are the most informative sentences in $\mathscr{L}_n$ (there are 2^n such sentences); intuitively, they are the descriptions of the possible worlds expressible within the language.

There is only one *true* constituent, denoted $C_\star$, which is the most informative true description of the actual world within $\mathscr{L}_n$. In this sense, $C_\star$ represents "the (whole) truth" about the world as expressible in $\mathscr{L}_n$. As c-theory A is compared with $C_\star$, its degree of verisimilitude can be simply defined in terms of the number of its true and false basic sentences, i.e., conjuncts that A shares or not, respectively, with $C_\star$. While this definition is restricted to the rather special case of conjunctive theories,

[1] After Popper's own definition of verisimilitude was shown to be untenable, authors like Oddie (1986), Niiniluoto (1987), Kuipers (2000), Schurz and Weingartner (2010), and Zamora Bonilla (1996) developed a number of post-Popperian theories of verisimilitude; see Niiniluoto (1998) for an historical survey, and Cevolani and Tambolo (2013) for an introduction to the verisimilitudinarian approach to scientific progress.

it will be sufficient for my purposes, since, as explained in Sect. 3, also the theory of dialectical structures is limited to such kind of theories.

Let us call each true conjunct of c-theory A a *match* of A, and each false conjunct a *mistake* of A. Moreover, let t_A, and f_A denote, respectively, the number of matches and the number of mistakes of A. Finally, let define the "degree of true (basic) content" $cont_t(A, C_\star)$ and "degree of false (basic) content" $cont_f(A, C_\star)$ of A as the normalized number of its matches and mistakes, respectively:

$$cont_t(A, C_\star) \stackrel{df}{=} \frac{t_A}{n} \quad \text{and} \quad cont_f(A, C_\star) \stackrel{df}{=} \frac{f_A}{n} \tag{1}$$

As said before, A is (highly) verisimilar if A makes many matches and few mistakes about $C_\star$. This intuition is captured by the following "contrast measure" of the verisimilitude of A (Cevolani et al. 2011, p. 188):

$$Vs_\phi(A) \stackrel{df}{=} cont_t(A, C_\star) - \phi cont_f(A, C_\star) \tag{2}$$

where $\phi > 0$.[2] Intuitively, different values of ϕ express the relative weight assigned to truth and falsity: the greater ϕ, the farther from the truth A will be due to its mistakes.

It is important to appreciate the difference between verisimilitude and other related notions. For instance, the (degree of) "accuracy" or "approximate truth" $acc(A)$ of c-theory A is arguably definable as the number of matches of A divided by the total number m_A of its conjuncts (cf. also Cevolani 2014a):

$$acc(A) \stackrel{df}{=} \frac{t_A}{m_A} \tag{3}$$

Note that, for any true c-theory A, $acc(A) = 1$. This fact alone shows that accuracy is only one "ingredient" of verisimilitude: A may be highly accurate (for instance because it is true) without being highly verisimilar. As an example, if $C_\star \equiv a_1 \wedge \cdots \wedge a_n$, then c-theories $A = a_1$ and $B = a_1 \wedge a_2$ are equally accurate, since they are both true and hence $acc(A) = acc(B) = 1$, but B is more verisimilar than A, since it conveys more true information about the world. For the same reason, the *false* c-theory $B' = \neg a_1 \wedge a_2 \wedge a_3 \wedge \cdots \wedge a_n$ may well be more verisimilar, although less accurate, than A. In Popper's words, verisimilitude "represents the idea of approaching comprehensive truth. It thus combines truth and content" (Popper 1963, p. 237): it follows that informative falsehoods may be more valuable, in terms of verisimilitude, than uninformative truths (e.g., tautologies). Among truths, however, verisimilitude covaries with logical strength: in the above example, since A and B are both true, and B entails A, then B is more verisimilar

[2] One may note that measure Vs_ϕ is not normalized, and varies between $-\phi$ and 1. A normalized measure of the verisimilitude of A is $(Vs_\phi(A) + \phi)/(1 + \phi)$, which varies between 0 and 1.

than A (cf. Niiniluoto 1987, sec. 6.6 for a discussion of this and other properties of verisimilitude measures).

3 Debate Dynamics: Evolving Dialectical Structures

The theory of dialectical structures (Betz 2013) aims at reconstructing multi-agent debates and studying, via computer simulations, the role of controversial argumentation in consensus formation and truth approximation. A debate is characterized by the "proponents" (agents) engaged in a given controversy, by a "pool" of relevant sentences, and by the arguments recognized by the participants in the debate. At any stage of the debate, each proponent adopts a given "position" (theory) concerning the sentences in question; as arguments are introduced in the debate, agents may be forced to change their positions, since these may become untenable in view of the proposed arguments.

With reference to the formal framework introduced in Sect. 2, the theory of dialectical structures can be presented as follows (cf. Betz 2013, ch. 2). The pool of sentences under discussion is the set of basic sentences (literals) of $\mathscr{L}_n$. At any moment of time, each participant adopts a position represented by a c-theory of $\mathscr{L}_n$: "complete" positions correspond to constituents (conjunctions of n literals) and "partial" positions to "proper" c-theories (conjunctions of m literals, with $0 < m < n$). So far, the study is limited to agents adopting complete positions or constituents as their theories (Betz 2013, p. 10; but see Sect. 5 below).

An argument is a deductively valid inference from some premises to a conclusion; all sentences in an argument are drawn from the pool of basic sentences. Thus, as arguments are proposed, previously unseen inferential relationships among the basic sentences of the underlying language are established.[3] The arguments in a debate are dialectically interconnected in the sense that each of them can "support" or "attack" other arguments.[4] More precisely, an argument attacks another argument when the conclusion of the former contradicts a premise of the latter; and an argument supports another argument when the conclusion of the former is equivalent to a premise of the latter. A *dialectical structure* represents a "snapshot" of the current state of the debate, displaying the set of arguments and the corresponding support and attack relations. For example, in the simple dialectical structure

[3]This implies, as noted by an anonymous referee, that the basic sentences appearing as premises or conclusions of some argument do acquire a richer "internal structure" to account for the inferential relation captured by the argument itself. At the present stage of Betz's theory, however, such internal structure plays no role, and both premises and conclusions are treated as unanalyzed basic sentences, especially as far as the definition of the verisimilitude proponents' positions is concerned.

[4]There are also other possible relationships among arguments (cf. Betz 2013, sec. 1.3), that however play no role in what follows. Even the support and attack relations are introduced here just to illustrate Betz's basic framework.

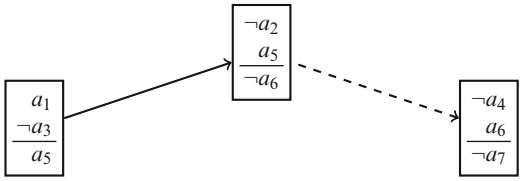


Fig. 1 A (very simple) dialectical structure; *continuous* and *dashed arrows* indicate the support and attack relation, respectively

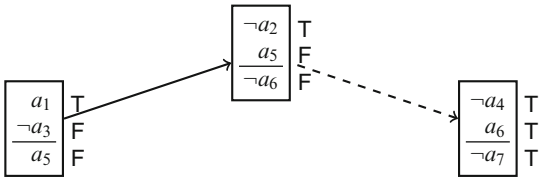


Fig. 2 A coherent complete position on the dialectical structure of Fig. 1. A "T" or "F" placed on the right of a sentence means that the corresponding sentence is, respectively, true or false

displayed in Fig. 1, the first argument (on the left) supports the second (middle) argument, which attacks the third argument (on the right).

At each step of a controversy, a proponent's position is "dialectically coherent" if it respects the argumentative structure of the debate, i.e., the inferential relationships so far established among the basic sentences of $\mathscr{L}_n$. More precisely, a complete position is dialectically coherent when, for any argument in a given structure, if its premises are true according to the position, also the conclusion is deemed true. As an example, Fig. 2 displays a coherent complete position—corresponding to constituent $a_1 \wedge \neg a_2 \wedge a_3 \wedge \neg a_4 \wedge \neg a_5 \wedge a_6 \wedge \neg a_7$ in $\mathscr{L}_7$—on the dialectical structure of Fig. 1.

How is belief dynamics modeled within the theory of dialectical structures? The central idea is that a debate is driven by the arguments which are proposed in the course of a controversy.[5] For instance, suppose that, with reference to the above dialectical structure, an argument is introduced to the effect that a_1 and a_6 jointly entail a_2. Then, the position displayed in Fig. 2 becomes incoherent: in fact, the proponent accepts both premises a_1 and a_6 but rejects the conclusion a_2 (since $\neg a_2$ is true according to the displayed position). Thus, the position has to be changed, on pain of inconsistency. Position dynamics is governed by an instance of the principle of conservatism or informational economy (Gärdenfors 1988): the proponent adopting a position which is made incoherent will shift to one of the *closest* coherent positions available. Such closest coherent positions are represented by the constituents which minimize the normalized Hamming distance from (the constituent representing) the current position (Betz 2013, p. 39). The Hamming distance between two constituents C_i and C_j is the number of atomic sentences on which the two constituents disagree; the normalized Hamming distance $\Delta(C_i, C_j)$

[5]The simplest method of introducing arguments in the simulation of a debate is randomly selecting some literals as premises and another as the conclusion; a good part of the theory, however, is devoted to studying more complex, and more interesting, argument construction mechanisms (see Betz 2013, pp. 8 ff. for a overview).

equals this number divided by n, the total number of atomic sentences. Since $\Delta(C_i, C_j)$ varies between 0 and 1, the (Hamming) similarity or closeness $s(C_i, C_j)$ between C_i and C_j can be defined as:

$$s(C_i, C_j) \stackrel{df}{=} 1 - \Delta(C_i, C_j) \tag{4}$$

i.e., as the normalized number of matches between C_i and C_j. Betz (2013, p. 39) takes $s(C_i, C_j)$ to measure the agreement between the two complete positions represented by C_i and C_j. Moreover, if $C_\star$ represents the true position, then $s(C, C_\star)$ is construed as the degree of verisimilitude of position C. As Betz (2013, p. 40, fn. 7) notes, this is an extreme special case of the Vs_ϕ measure of verisimilitude introduced in Sect. 2, since it is easy to prove that for any constituent C[6]:

$$\text{for } \phi = 0, \text{ then } Vs_0(C) = cont_t(C, C_\star) = s(C, C_\star) \tag{5}$$

Finally, Betz (2013, p. 40) introduces the idea of "debate-wide mean verisimilitude". This is simply defined as the average degree of verisimilitude over all individual positions—a value which intuitively expresses how close to the truth is the group of proponents taken as a whole. This mean verisimilitude value is used by Betz to assess the truth conduciveness of different epistemic practices as dialectical structures evolve during a debate (Betz 2013, part II).

4 Truth Approximation in Debate Dynamics

When does belief dynamics lead a group of epistemic agents closer to the truth, i.e., increases the verisimilitude of their individual and/or aggregated theories? As recent work has shown, it is very hard to specify suitably general conditions under which revising or merging individual beliefs effectively tracks truth approximation.[7] The same is true in the present case, since the introduction of new valid arguments in a debate can both increase and decrease, depending on the specific circumstances, the verisimilitude of the individual positions, and hence the debate-wide mean verisimilitude. The following example is a case in point.

Example 1. Let us consider a simple pool of sentences, given by the eight basic sentences of $\mathscr{L}_4$: $a_1, \neg a_1, \ldots, a_4, \neg a_4$. Suppose that, at the initial stage of some debate, three agents adopt the following complete positions:

[6]More generally, if C is a constituent then $Vs_\phi(C)$ only depends on its degree of truth content $cont_t(C, C_\star)$, i.e., on $s(C, C_\star)$. In fact, since for constituents $cont_t(C, C_\star) = 1 - cont_f(C, C_\star)$, then $Vs_\phi(C) = cont_t(C, C_\star) - \phi cont_f(C, C_\star) = (\phi + 1)cont_t(C, C_\star) - \phi$.

[7]See, in particular, Niiniluoto (1999, 2011) and the papers collected in Kuipers and Schurz (2011); for the relations between the basic feature approach and different accounts of belief dynamics see Cevolani et al. (2011, 2013), and Cevolani (2013, 2014b).

$$
\begin{aligned}
X: \ & a_1 \ \wedge \ a_2 \ \wedge \neg a_3 \wedge \ a_4 \\
Y: \ & a_1 \ \wedge \neg a_2 \wedge \ a_3 \ \wedge \neg a_4 \\
Z: \ & \neg a_1 \wedge \ a_2 \ \wedge \neg a_3 \wedge \neg a_4
\end{aligned}
$$

Assuming that $a_1 \wedge a_2 \wedge a_3 \wedge a_4$ is the true position, is it easy to calculate that $Vs_0(X) = \frac{3}{4}$, $Vs_0(Y) = \frac{1}{2}$, and $Vs_0(Z) = \frac{1}{4}$, and that the debate-wide mean verisimilitude is $\frac{1}{2}$. Now suppose that an argument is introduced in the debate to the effect that a_1 and a_3 jointly entail a_4. While positions X and Z remain coherent, position Y is now dialectically incoherent, since the agent adopting Y deems both premises a_1 and a_3 true, but the conclusion a_4 false. As a consequence, position Y has to be changed to one of the closest positions, i.e., to one of the following constituents:

$$
\begin{aligned}
Y': \ & a_1 \ \wedge \neg a_2 \wedge \neg a_3 \wedge \neg a_4 \\
Y'': \ & \neg a_1 \wedge \neg a_2 \wedge \ a_3 \ \wedge \neg a_4 \\
Y''': \ & a_1 \ \wedge \neg a_2 \wedge \ a_3 \ \wedge \ a_4
\end{aligned}
$$

As one can check, $Vs_0(Y') = Vs_0(Y'') = \frac{1}{4}$ and $Vs_0(Y''') = \frac{3}{4}$. In two cases out of three, corresponding to the choice of either Y' or Y'' as the new position, this will be less verisimilar than the old position Y; accordingly, the introduction of the argument above will decrease the debate-wide mean verisimilitude (from 0.5 to about 0.42). Only in the third case, corresponding to the change from Y to Y''', does the individual verisimilitude and hence the debate-wide mean verisimilitude increase (the latter from 0.5 to about 0.58).

It is perhaps worth noting that, in the above example, the newly introduced argument is not only valid but also sound, in the sense that both premises are true. Thus, even accepting an inference from true premises to a true conclusion may actually lead proponents farther from the truth about the matter under discussion.

Cases like Example 1 show that debate dynamics does not track truth approximation in general, and can sometimes hinder, in fact, cognitive progress. In more specific situations, however, there may be other factors that counterbalance the negative effects of cases like these. For instance, the agents may already share some *true* information about the world, construed as common evidence or background knowledge (Betz 2013, p. 14 and ch. 12). This is modeled in Betz's framework by fixing, right at the beginning of a simulation, the truth values of some of the basic sentences to the correct value, hence helping the group both to attain consensus and to approach the truth about the domain. In any case, by repeatedly simulating a high number of abstract debates (evolving dialectical structures as those introduced in Sect. 3), one can identify some general patterns of how argumentation influences the group capability of approaching truth. In particular, Betz's results tend to confirm the idea that, under suitably specified circumstances, controversial argumentation is instrumental both for attaining a consensus and for approaching truth; moreover,

consensus formation is a good, although fallible, indicator of truth approximation.[8] Without aiming at an even cursory presentation of these results (for which see Betz 2013, secs. 1.4–1.5), in the following section I focus on the link between the theory of dialectical structures and the notion of verisimilitude as introduced in Sect. 2.

5 Extending the Theory of Dialectical Structures

In the analysis of debate dynamics presented so far, I only considered agents maintaining complete positions (constituents). This means that all participants in a debate have a definite opinion on each element of the relevant pool of basic sentences. As Betz (2013, p. 10) notes, this is a quite limiting assumption: simulating debates among agents who adopt *partial* positions would allow for a more realistic representation of actual controversies, and may shed new light on the truth approximation issue.

An agent adopting a partial position (i.e., a proper c-theory) suspends the judgment on at least some of the basic sentences in the pool. As an example, Fig. 3 displays the partial position corresponding to c-theory $a_1 \wedge \neg a_2 \wedge \neg a_3 \wedge a_6$ in $\mathscr{L}_7$. Note that this position is (dialectically) incoherent, since it cannot be "extended" to a coherent complete position (Betz 2013, pp. 35–36). In fact, since according to the agent a_1, $\neg a_2$, $\neg a_3$ and a_6 are all true, given the first argument (on the left) also a_5 should be true; but since the second argument (in the middle) has a false conclusion ($\neg a_6$) and a true premise ($\neg a_2$), the remaining premise (a_5) should be false. Since the dialectical structure forces the proponent to assign opposite truth values to the same sentence (a_5), this position is incoherent.

Allowing for agents adopting partial positions requires two relevant modifications of the theory of dialectical structures. First, the mechanism governing belief dynamics, which guarantees that the agents' positions remain coherent in the course of the debate, has to be extended. Second, the verisimilitude of partial positions has to be adequately defined.

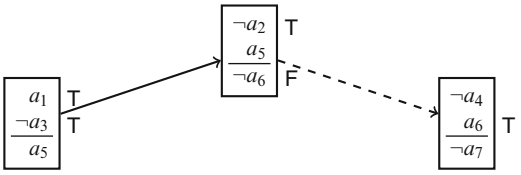


Fig. 3 An incoherent partial position on the dialectical structure of Fig. 1

[8] See Betz (2013, chs. 3 and 10) for a presentation of these results and of the underlying assumptions. Given space limitations, I cannot discuss the interesting relationships between truth approximation and consensus formation. As Betz (2013, pp. 13 ff.) notes, there is a clear asymmetry between the two: in fact, reaching the truth implies having reached a consensus, while the converse does not hold in general: cases of "spurious consensus"—i.e., agreement on some false position—are very well possible both in scientific and ordinary controversies.

As far as the former problem is concerned, the example in Fig. 3 makes clear that a proponent adopting an incoherent partial position has many different options to make it coherent. In fact, while a complete position can only be modified by replacing one or more of the conjuncts of the corresponding constituent with their negation (cf. Example 1), an agent adopting a partial position can also withhold judgment on a previously accepted sentence to restore coherence. As an example, the position in Fig. 3 can be made coherent by replacing, say, $\neg a_2$ with its negation (thus shifting to the new position $a_1 \wedge a_2 \wedge \neg a_3 \wedge a_6$), but also by suspending the judgment, e.g., on a_1 (thus weakening the agent's position to $\neg a_2 \wedge \neg a_3 \wedge a_6$); and indeed in many other ways. This raises the problem of how the new, coherent position should be selected. Since the Hamming distance is only defined for constituents (cf. Eq. 4), the distance minimization method introduced in Sect. 3 can not be applied in this case. In short, one needs an extended definition of the distance (or, equivalently, closeness) between any two proper c-theories in order to identify the closest partial positions to a given, incoherent partial position.[9]

Concerning the definition of the degree of verisimilitude of a partial position A, Betz (2013, pp. 39–40, especially fn. 7) suggests to use the number t_A of the matches (true basic sentences) of A divided by the total number m_A of the sentences in A. This proposal amounts to defining the verisimilitude of a c-theory A as its accuracy or approximate truth $acc(A)$ (cf. Eq. 3 in Sect. 2). It is worth noting that, as far as complete positions (i.e., constituents) are concerned, this definition indeed agrees with the one given in Sect. 3 in terms of the Hamming distance (see Eq. 4). In fact, it is easy to check that, for any constituent C:

$$acc(C) = cont_t(C, C_\star) = \frac{t_C}{n} = Vs_0(C) \tag{6}$$

In other words, the accuracy of a constituent equals its degree of true content, which is taken by Betz to express the verisimilitude of the corresponding complete position. On the other hand, this can be accepted as an adequate definition of verisimilitude only because, for constituents, verisimilitude and accuracy covary: since any two constituents are equally (and maximally) informative, the greater the accuracy, the greater the verisimilitude (cf. also Eq. 5).

This strict relation between verisimilitude and accuracy, however, does not hold when proper c-theories are considered (cf. Sect. 2). This implies, among other things, that two agents may adopt two different highly accurate (or even true) c-

[9] Such an extended definition is also needed if other notions employed by Betz have to be generalized to agents adopting partial positions. These include, for instance, the agreement between two positions, the debate-wide mean normalized agreement, and the resulting notion of consensus (see Betz 2013, pp. 39–40). In this connection, the analysis of "theory distance" developed by verisimilitude scholars (see, e.g., Niiniluoto 1987, sec. 6.7 and the corresponding references) may provide valuable suggestions for such a definition. In particular, the feature contrast measures of verisimilitude introduced in Sect. 2 are arguably applicable to a generalized definition of the similarity between c-theories. The discussion of such a generalization, however, has to be left to another occasion.

theories as their partial positions, but still these may be very far from the truth and also far from each other. As an example, again assuming that $C_\star \equiv a_1 \wedge \cdots \wedge a_n$ is the truth, suppose that the two agents adopt $A = a_1 \wedge a_2$ and $B = a_3 \wedge a_4$, respectively, as their partial positions. Then, the accuracy of their positions is maximal ($acc(A) = acc(B) = 1$) but their verisimilitude may well be low, if n is sufficiently high.

In conclusion, in order to adequately generalize the theory of dialectical structures to agents adopting partial positions, one needs to introduce an appropriate measure of their verisimilitude. To this purpose, the feature contrast measure Vs_ϕ defined in Sect. 2 suggests itself, being equally applicable to proper c-theories and to constituents. Allowing for agents to maintain partial positions, and using a measure like Vs_ϕ to assess their verisimilitude, would predictably lead to a richer and more realistic account of truth conduciveness within debate dynamics. It is admittedly difficult, however, to anticipate what implications this extension would have for Betz's results concerning truth approximation and consensus formation. Indeed, it is well possible that the introduction of partial positions makes debate dynamics too complex to replicate the positive tendency toward reciprocal agreement and truth displayed by Betz's simulations. Interesting as it may be, an exploration of how this extension would affect the results of the corresponding simulations exceeds the limits of the present paper.

6 Conclusions: Truth Tracking vs. Truth Approximation

Starting at least with the work of Keith Lehrer in the 1970s of the past century (Lehrer and Wagner 1981), philosophers of science have developed a number of models of belief dynamics and consensus formation in order to evaluate the truth conduciveness of different social practices. Examples are Kitcher's account of consensus practices and of the division of cognitive labor (Kitcher 1993, ch. 8) and the simulation-based approaches to so called opinion dynamics (Hegselmann 2004; Lehrer and Wagner 1981). Other relevant approaches are the theory of judgment aggregation (List 2012), developed at the interface between political theory and social choice theory (see also Zamora Bonilla 2007, for a philosophy of science perspective), and the theories of belief revision (Gärdenfors 1988) and of belief merging (Konieczny and Pino Pérez 2011) as studied in logic and AI.

In an interesting comparative survey, Betz (2013, sec. 1.7) argues that these approaches can be classified according to "the degree of logical competence which agents are assumed to possess according to the corresponding approach" (*ibidem*, p. 25). Some accounts (like standard belief revision theory) assume that epistemic agents are "logically omniscient", in the sense that their beliefs form consistent and logically closed sets with a complex internal inferential structure. Other accounts, on the contrary, depict their agents as "logically ignorant" subjects (Betz 2013, p. 27): these are exemplified by the Lehrer-Wagner and the Hegselmann-Krause models of opinion dynamics, where agents' beliefs are represented as point estimates of a real-valued parameter, with no inferential structure whatsoever (but see Riegler and Douven 2009).

As valuable as it may be, Betz's competence-based classification is better supplemented, for our present purposes, with a conceptually independent distinction concerning the issue of truth conduciveness. In Sect. 2, I introduced the notion of verisimilitude as characterizing the main epistemic goal of (a community of) rational truth-seeking agents. Such an approach naturally leads to what was called the *truth approximation* issue: i.e., exploring whether, and under what conditions, revising and merging the beliefs of a group of agents lead them closer to the truth about the relevant domain or, which is the same, increases the verisimilitude of their positions. A different question is the following: under what assumptions concerning the reliability of the agents in the group and the merging procedure, the whole truth about the domain (i.e., $C_\star$) is eventually identified? This is known in the literature as the *truth tracking* issue. The analysis of this problem was partly motivated by the so called Condorcet's jury theorem in political science, which roughly says that if the agents are sufficiently reliable—i.e., their individual probability of correctly judging the truth or falsity of a proposition is greater than 0.5—and form their opinions independently, then the probability that the majority of the agents correctly judges a proposition approaches 1 as the number of agents tends to infinity. This issue has been variously explored, by means of computer simulations, in a number of areas, including the logical theory of belief merging (Konieczny and Pino Pérez 2011), the Hegselmann-Krause model of opinion dynamics (Douven and Kelp 2011), and the theory of judgment aggregation (Hartmann and Sprenger 2012).

Truth approximation, on the one hand, and truth tracking, on the other, can be construed as different, but not incompatible, views of the truth conduciveness issue. In both cases the problem is how to gradually approach to the whole truth about the target domain, but the approach differs under a number of aspects. Without aiming at more than a partial and revisable characterization, the following main differences can be emphasized. When truth approximation is at issue,

(i) the focus is on the beliefs or theories of one or a few agents (with few exception, verisimilitude and related matters have been mainly studied, so far, within "traditional" or "individualistic" epistemology);
(ii) these beliefs are typically represented by incomplete or partial theories (like proper c-theories or partial positions) in some language, the issue being how to "complete" such theories in order to make them closer to the whole truth;
(iii) accordingly, the focus is on the *micro*dynamics of belief, and the revision or aggregating mechanism is usually sophisticated (like AGM revision rules, merging procedures, or the maximization of expected verisimilitude);
(iv) finally, this microdynamics is studied analytically, for instance by proving theorems showing how the verisimilitude of individual or collective theories is increased or maximized.

In comparison, the truth tracking issue raises different concerns:

(i) the focus is on the group or community which the epistemic agents belong to (the epistemological perspective is fully "social");

(ii) the agents' beliefs are typically represented by complete theories (like constituents or complete positions) and one then studies how such complete theories spread in the community, which may eventually converge toward the true complete position;
(iii) the *macro*dynamics of such evolution is the main concern, while the revision/aggregating procedure is comparatively less sophisticated (like the imitation of peers' average opinion in the Lehrer-Wagner and Hegselmann-Krause models);
(iv) due to the complex interaction among the beliefs of many different agents, computer simulations are usually the only available means to study the general truth tracking effectiveness of different procedures.

In the light of Betz's distinction between logically omniscient and logically ignorant approaches, theories of truth approximation tend to be associated with the former, and theories of truth tracking with the latter.

As the above comparison should make clear, truth approximation and truth tracking are strongly related ways of studying relevant problems in veristic social epistemology. Depending on the specific problem and on other circumstances, one approach may prove more useful than the other; more generally, each is characterized by strong and weak points. For instance, simulation-based accounts of truth tracking often enlighten the dynamics of entire communities of epistemic agents, but tend to obscure the relevant microdynamics at the level of individual theories; on the other hand, analyses of truth approximation often rely on a detailed account of individual belief change but quickly run into troubles with complex multi-agents interactions. As a mid-way approach combining aspects of both kinds of accounts, the theory of debate dynamics, especially when extended along the lines suggested in Sect. 5, seems able to provide valuable insights on both the truth approximation and the truth tracking issue.

Acknowledgements I would like to thank Gregor Betz for useful discussion on the topics of the paper, and Luca Tambolo and two anonymous referees for very detailed comments and corrections on an earlier draft of the paper. Financial support from the priority program *New Frameworks of Rationality*, SPP 1516 (Deutsche Forshungsgemeinshaft, grant CR 409/1-2), and from the FIRB project *Structures and Dynamics of Knowledge and Cognition* (Italian Ministry of Scientific Research, Turin unit, D11J12000470001) is gratefully acknowledged.

References

Betz G (2013) Debate dynamics: how controversy improves our beliefs. Springer, Dordrecht
Cevolani G (2013) Truth approximation via abductive belief change. Logic J IGPL 21(6):999–1016
Cevolani G (2014a) Strongly semantic information as information about the truth. In: Ciuni R, Wansing H, Willkommen C (eds) Recent trends in philosophical logic. Springer, Cham, pp 59–74
Cevolani G (2014b) Truth approximation, belief merging, and peer disagreement. Synthese 191(11):2383–2401

Cevolani G, Tambolo L (2013) Progress as approximation to the truth: a defence of the verisimilitudinarian approach. Erkenntnis 78(4):921–935
Cevolani G, Crupi V, Festa R (2011) Verisimilitude and belief change for conjunctive theories. Erkenntnis 75(2):183–202
Cevolani G, Festa R, Kuipers TAF (2013) Verisimilitude and belief change for nomic conjunctive theories. Synthese 190(16):3307–3324
Douven I, Kelp C (2011) Truth approximation, social epistemology, and opinion dynamics. Erkenntnis 75:271–283
Gärdenfors P (1988) Knowledge in flux: modeling the dynamics of epistemic states. MIT, Cambridge
Goldman A (1999) Knowledge in a social world. Oxford University Press, Oxford
Goldman A (2010) Social epistemology. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, summer 2010 edn
Hartmann S, Sprenger J (2012) Judgment aggregation and the problem of tracking the truth. Synthese 187(1):209–221
Hegselmann R (2004) Opinion dynamics – insights by radically simplifying models. In: Gillies D (ed) Laws and models in science. King's College Publications, London, pp 19–46
Kitcher P (1993) The advancement of science: science without legend, objectivity without illusions. Oxford University Press, New York
Konieczny S, Pino Pérez R (2011) Logic based merging. J Philos Logic 40:239–270
Kuipers TAF (2000) From instrumentalism to constructive realism. Kluwer Academic, Dordrecht
Kuipers T, Schurz G (eds) (2011) Special issue on "Belief revision aiming at truth approximation". Erkenntnis 75(2):151–283
Lehrer K, Wagner C (1981) Rational consensus in science and society. Reidel, Boston
List C (2012) The theory of judgment aggregation: an introductory review. Synthese 187(1):179–207
Niiniluoto I (1987) Truthlikeness. Reidel, Dordrecht
Niiniluoto I (1998) Verisimilitude: the third period. Br J Philos Sci 49(1):1–29
Niiniluoto I (1999) Belief revision and truthlikeness. In: Hansson B, Halldén S, Sahlin NE, Rabinowicz W (eds) Internet Festschrift for Peter Gärdenfors, Department of Philosophy, Lund University, Lund. http://www.lucs.lu.se/spinning/
Niiniluoto I (2011) Revising beliefs towards the truth. Erkenntnis 75(2):165–181
Oddie G (1986) Likeness to truth. Reidel, Dordrecht
Popper KR (1963) Conjectures and refutations: the growth of scientific knowledge, 3rd edn. Routledge and Kegan Paul, London
Riegler A, Douven I (2009) Extending the Hegselmann-Krause model III: from single beliefs to complex belief states. Episteme 6(2):145–163
Schurz G, Weingartner P (2010) Zwart and Franssen's impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude. Synthese 172:415–436
Zamora Bonilla J (1996) Verisimilitude, structuralism, and scientific progress. Erkenntnis 44:25–47
Zamora Bonilla J (2007) Optimal judgment aggregation. Philos Sci 74(5):813–824

Wise Crowds, Clever Meta-Inductivists

Paul D. Thorn

1 Introduction

In several recent papers (2008, 2009b), Gerhard Schurz proposed a response to Hume's problem of induction, based on *meta-induction*. In its various forms, meta-induction proceeds by considering the past track record of other agents (and/or prediction methods), and makes predictions of future events by reasoning that agents (and/or prediction methods) that have been successful in the past will be successful in the future. Schurz demonstrated that, under plausible conditions, various forms of meta-induction are guaranteed to yield optimal results, in the sense of having predictive success rates that converge to the success rate of the meta-inductivist's most successful competitor.

The optimality of meta-induction *appears* to provide a strong prescription for would-be predictors. But the matter is, perhaps, not so simple. The core injunction of meta-induction is to *copy* the strategies and predictions of those individuals who have proven most reliable. The prescriptions of meta-induction are thus in tension with prescriptions implicit in recent formal and empirical work on the Wisdom of Crowds. Such work emphasizes the importance of agents making their predictions (and judgments) *independently* of the predictions of other agents.

Francis Galton's account of a contest that occurred at the 1906 West England Fat Stock and Chicken Exhibition is a popular touchstone for discussions of the Wisdom of Crowds, and serves as compelling, if anecdotal, illustration of the 'wise crowd effect'. In the contest recounted by Galton, attendees at a livestock exhibition could observe a mature ox, and had the opportunity to guess its weight. Seven hundred

P.D. Thorn (✉)
Philosophy Department, University of Duesseldorf, Universitaetsstr. 1,
Duesseldorf 40204, Germany
e-mail: thorn@phil-fak.uni-duesseldorf.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_6

and eighty seven persons entered the contest, and offered wide-ranging guesses. The remarkable fact about these guesses resided in their average, the 'judgment of the crowd'. The crowd guessed that the ox would weigh 1197 lb, while the ox weighed 1198 lb.

Empirical studies have illustrated that the judgments of crowds (i.e., the average value of the judgments of a group's members) are remarkably reliable in the face of certain types of query (Surowiecki 2004, 5–22; Page 2007, 178). It is also straightforward to construct formal models of individual judgment wherein the average value of the judgments of a group tends to be very accurate. Recent empirical studies also show that the accuracy of a crowd's judgment can be compromised when agents within the group are aware of the judgments made by other group members (and are thus able to *imitate* other group members) (Lorenza et al. 2011). Similarly, well known formal models of 'wise crowds' require that the judgments of a group's members be *stochastically independent* of the judgments of other members of the group (conditional on the value of the predicted event). So select empirical and formal results *suggest* that imitating the judgments of other group members is, *contra* meta-induction, a bad thing.

In a recent paper, Thorn and Schurz (2012) presented results concerning the impact on a group's performance that may result from having members of a group adopt meta-inductive methods (cf. Schurz 2012). In this paper, I replicate a selection of those results, illustrating that, in a variety of circumstances, the adoption of meta-inductive methods can decrease the accuracy of the aggregate judgment of the group. I then expand on previous work by considering three simple measures by which meta-inductive prediction methods may improve their own performance, while simultaneously mitigating their negative impact on the aggregate judgment of the group.

2 The Optimality of Global Meta-Induction

To demonstrate the optimality of meta-induction, Schurz (2008, 2009b) introduced the notion of a *prediction game*, consisting of:

1. An infinite sequence $(e) = (e_1, e_2, \dots)$ of events, whose values are drawn from the unit interval, i.e., $e_n \in [0,1]$, for each round, n, of the game (and from $\{0,1\}$ in the case of *binary* prediction games).
2. A finite set of players, Π, whose task in each round is to predict the value of the next event. "$p_n(P)$" denotes the prediction of *player* P at time n, which is delivered at time $n-1$. The players in Π include: one or several meta-inductivist players of various kinds (see below), and a finite set of non-MI-players players $P_1, \dots, P_m$. It is assumed that the MI-players make their predictions after the non-MI-players, and may thus imitate the predictions of the non-MI-players.

Within prediction games, the deviation of a prediction p_n from the event e_n is measured by a normalized loss function $l(p_n, e_n) \in [0,1]$. A prominent loss-

function measures the absolute difference between event and prediction, $|e_n - p_n|$, but the optimality theorems described below are not restricted to this loss function: Theorem 1 holds for *monotonic* loss-functions, and Theorem 2 holds for *convex* loss-functions. The *score*, $s(p_n, e_n)$, obtained in round n is defined as $1 - l(p_n, e_n)$. The *success rate*, $suc_n(P)$, of player P, at time n, is $\Sigma_{1 \leq i \leq n}\, s(p_i(P), e_i)/n$. Finally, $maxsuc_n$ is the maximal success rate of the non-MI-players at time n.

The simplest type of meta-induction is called "imitate-the-best". In each round, bMIs (players who employ imitate-the-best meta-induction) imitate the prediction of the non-MI-player with the so-far highest success rate. bMIs change their favorite player as soon as another player achieves a higher success-rate. If there are several best players, bMIs chooses her favorite by a predefined ordering of the non-MI-players. The central result concerning imitate-the-best prediction method is as follows:

Theorem 1 (Schurz 2008) For every prediction game $((e),\{P_1, \ldots, P_m, bMI\})$ that contains a *best* non-MI-player, B, after some round n_B (i.e., $suc_n(B) > suc_n(P_i)$ for all $n \geq n_B$ and $P_i \neq B$), the following holds:

(1.1) *Short run*: For all rounds n, $suc_n(bMI) \geq maxsuc_n - (n_B/n)$.
(1.2) *Long run*: As n approaches ∞, $suc_n(bMI)$ converges to $maxsuc_n$.

The assumption of Theorem 1 that there is a *best* non-MI-player, B, after some finite number of rounds is rather strong. A satisfactory solution to Hume's problem calls for a meta-inductive strategy whose performance is optimal when this assumption does not hold. *Weighted meta-induction* fills this role. wMIs (players who employ weighted meta-induction) predict a weighted average of the predictions of the so-far 'most attractive' players. The attractivity $at_n(P)$ of player P at time n is P's surplus success-rate compared to the wMI's success: $at_n(P) = suc_n(P) - suc_n(wMI)$, provided $suc_n(P) > suc_n(wMI)$, otherwise $at_n(P) = 0$. A wMI's predictions are defined as $p_{n+1}(wMI) = \sum_P(at_n(P) \cdot p_{n+1}(P))/\sum_P(at_n(P))$, where P ranges over all accessible players. (If no player has positive attractivity, the wMI makes a random guess.) The following establishes weighted meta-induction's long-run optimality:

Theorem 2 (Schurz 2008, cf. Cesa-Bianchi and Lugosi 2006) For every real-valued prediction game $((e),\{P_1, \ldots, P_m, wMI\})$ whose loss-function $l(p_n, e_n)$ is *convex* in the argument p_n, the following holds:

(2.1) *Short run*: $\forall n \geq 1$: $suc_n(wMI) \geq maxsuc_n - \sqrt{(m/n)}$.
(2.2) *Long-run*: As n approaches ∞, $suc_n(wMI)$ converges to $maxsuc_n$.

Theorem 2 does not apply directly to binary prediction games, because a wMI's predictions are real-valued. However, Theorem 2 can be generalized to binary valued predictions, by positing a *population* of sufficiently many, say k, meta-inductivists, who imitate the predictions of each attractive non-MI-player, P, with a population share that is approximately equal to P's attractivity. The *mean success rate* of such populations approximates the maximal success rate of the most attractive non-MI-players, with an additional maximal short-run loss of $1/(2 \cdot k)$

(Schurz 2008, 2009a, b). Similar convergence results hold for the *expected* success-rate a meta-inductivist who predicts respective outcomes with probability equal to the population shares represented among such groups.

3 The Wise Crowd

Following in the footsteps of some recent monographs (Surowiecki 2004; Page 2007, 179), I will say that the *judgment of a crowd* with respect to *query* is the average response of its members (rounded if necessary), treating affirmation as *one* and disaffirmation as *zero*, in the case of binary predictions. Along with the preceding convention, I say that *a crowd is wise* to the extent that its judgments are *accurate*.

Anecdotes such at Francis Galton's (Sect. 1) are relatively widespread, and it is clear that the judgment of a crowd with respect to some kinds of query frequently exhibits uncanny accuracy (i.e., the wise crowd effect). In this same vein, Jack Treynor illustrated the wise crowd effect by having groups of students guess the number of jelly beans contained in a large jar (Surowiecki 2004, 5). Various markets, from stock exchanges to professional football betting lines (Surowiecki 2004, 12–15), and *prediction markets* such as the Iowa Electronic Markets and the Hollywood Stock Exchange have demonstrated the accuracy of groups of independently acting individuals in making various kinds of prediction (Surowiecki 2004, 17–22; Page 2007, 178).

An early mathematical model that exhibits a sufficient condition for a wise crowd is described by the Condorcet Jury Theorem. The theorem considers a group of individuals, where for each member of the group the probability is r ($r > 0.5$) that that member of the group will make a correct judgment regarding the truth value of some proposition. It is further assumed that each individual's likelihood of making a correct judgment is *stochastically independent* of whether other members of the group make correct judgments. Under these conditions, the theorem tells us that the likelihood that the majority response of its members is correct converges to *one* as the size of the group approaches ∞.

The Weak Law of Large Numbers suggests an obvious means of modeling wise crowds in the case of real-valued events. Where ε is any number greater than 0, and X_n is a sample of n independent identically distributed random variables with mean μ, the Law of Large Numbers tells us that the probability that the mean value of the elements of X_n differs from μ by more than ε converges to *zero* as n approaches ∞. We may thus conceive of the sample X_n as a set of predictions made by a group of n individuals about the value some unknown quantity μ, where the elements of X_n are independently and identically distributed around μ. In that case, the Law of Large Numbers tells us, for all $\varepsilon > 0$, that the probability that the group's judgment about the value of μ differs from μ by more than ε goes to zero as n approaches ∞.

The Condorcet Jury Theorem and the Law of Large Numbers provide models describing how the average judgment of a group's members can be extremely accurate, provided the group is large and its members have some *truth-bias*, i.e., under the condition that there is a *better chance than not* that each group member makes a true judgment in the case of true/false queries, and under the condition that each group member's judgment is distributed around the true value with a mean value that is identical to the true value, in the case of real-valued queries.[1]

While the wise crowd effect recommends that forecasters make their predictions independently, the optimality of meta-induction suggests that forecasters should imitate the most successful forecasters whose predictions are accessible. So there is a tension between the preconditions for wise crowds, and the injunctions of meta-induction. In the face of this tension, Thorn and Schurz (2012) evaluated the impact on group performance that may result from having members of a group adopt meta-inductive methods. Their results illustrate a variety of conditions under which *replacing* non-imitative players by meta-inductivists reduces the accuracy of the aggregate judgment of the group. After introducing the formal framework of (Thorn and Schurz 2012), I summarize some of their results. I then consider three simple measures by which meta-inductive prediction methods may improve their own performance, while simultaneously mitigating their negative impact on group performance.

4 The Formal Setup

Departing slightly from the prediction games described in Sect. 2, the simulations described here include the following elements:

1. A quadratic grid consisting of $100 \times 100 = 10{,}000$ cells. Each cell corresponds to an individual player.
2. For some simulations, each agent has *access* to the success rates and the present judgment of every other player. In other simulations, each player only has access to information concerning the players in her *Moore-neighborhood,* i.e., to herself and the eight immediately surrounding players.
3. The event sequence is either: a random sequence of values chosen according to a uniform probability distribution on the unit-interval [0,1], or a binary event sequence generated by rounding the elements of a sequence of the preceding sort, where values greater than 0.5 are rounded to 1. In the case of a binary event sequence, players are required to predict that the true value of any event is 0 or 1. In the case of the real-valued event sequence, players may predict *any* real number (thereby permitting the possibility of arbitrarily large errors).

[1]The independence assumptions under which the two theorems apply limit the applicability of the corresponding models. A more realistic formal model is found in (Page 2007), drawing on work by Krogh and Vedelsby (1995).

4. In the case of a binary event sequence, each player has a predefined *independent reliability*, r, which is the player's probability of making a correct prediction in any given round (as determined by Bernoulli trials), assuming she bases her predictions solely on her own abilities, and independently of other players. Each player's *independent unreliability*, u, is $1 - r$. In the case of a real-valued sequence, each player's prediction is assumed to be normally distributed with a mean identical to the true event-value, where the mean absolute deviation is the player's *independent unreliability*, u.[2]
5. The game consists of rounds, but now in addition, each round consists of successive *cycles*, in which predictions may be updated by imitating the predictions of other accessible players.
6. In addition to their independent prediction abilities, some players apply one of the following imitative prediction methods to other *accessible* players:

   (a) Weighted meta-induction wMI[3]: In the face of a real-valued event sequence, wMIs predict the *attractivity* weighted average of the predictions of those players accessible to the wMI. In the case of binary event-sequences, wMIs predict the *rounded* attractivity weighted average of the predictions of those players accessible to the wMI. In face of both real-valued and binary event sequences, wMIs predict by independent means in the first round, in the first cycle of each round, and whenever they themselves have the highest success rate.
   (b) Peer-imitation: Peer-imitators predict an *unweighted* average of the predictions of those players accessible to the peer-imitator.

In contrast to the sort of prediction games described in Sect. 2, the present setup allows for mutual imitation between imitative players. Since a player can imitate another player only *after* that player has made a prediction, the imitation process is now modeled via successive *update cycles*, in which players may imitate the predictions that her favorite(s) delivered in the *previous* cycle. In the first cycle of each round, each player delivers a prediction based on her independent abilities. In all following cycles, independent players repeat their initial prediction, while imitative players apply their imitative prediction method to the predictions made by accessible players in the previous cycle. This continues until a preselected maximum number of cycles is reached. After the *final predictions* for a round are determined, the actual success rate for each player is updated, and a new round (with a new sequence of prediction cycles) begins, until the final round of the game is reached.

[2] So the standard deviation of an agent's independent guess is $u \cdot \sqrt{(2/\pi)}$, since $\sqrt{(2/\pi)}$ is the ratio of the mean absolute deviation to the standard deviation in the case of normal distributions.

[3] As space is limited, I focus exclusively of weighted meta-induction, which performed better than imitate-the-best meta-induction in the simulations studied in (Thorn and Schurz 2012).

5 Groups with Universal Accessibility

In the present section, I replicate results from (Thorn and Schurz 2012) in order to illustrate some effects which ensue when members of a group adopt weighted meta-induction, in cases where the predictions and success rates of all agents are accessible to all agents. Figure 1 presents results for binary predictions, comparing populations composed wholly of independent predictors to ones composed wholly of wMIs, in games that lasted 1000 rounds, with 2 cycles per round. The independent *unreliability* of the agents is the independent variable. The respective mean *error rate* (i.e., the mean linear distance between predicted and actual values) is the dependent variable.

Figure 2 presents results analogous to Fig. 1 for real-valued predictions. Note that for all the figures concerning real-valued predictions, the scale for the intervals [0,1], [1,5], and [5,20] differs in order to magnify small differences in the values of the dependent variable that occur particularly in the interval [0,1].

The results represented in Figs. 1 and 2 reflect the fact that the mean individual error rates of non-imitators converge to their independent unreliabilities. We also observe (as predicted by the Condorcet Jury Theorem and the Law of Large Numbers) that the mean group error for non-imitators is usually quite low. In the binary case, setting individual unreliability, u, to be somewhat less than 0.5 is sufficient to make it a practical certainty that the crowd is very wise (while setting u to be somewhat higher than 0.5 is sufficient to ensure that the crowd is very *unwise*). In the real-valued case, even a modest truth bias, such as $u = 20$, is sufficient to achieve a relatively low mean group error.

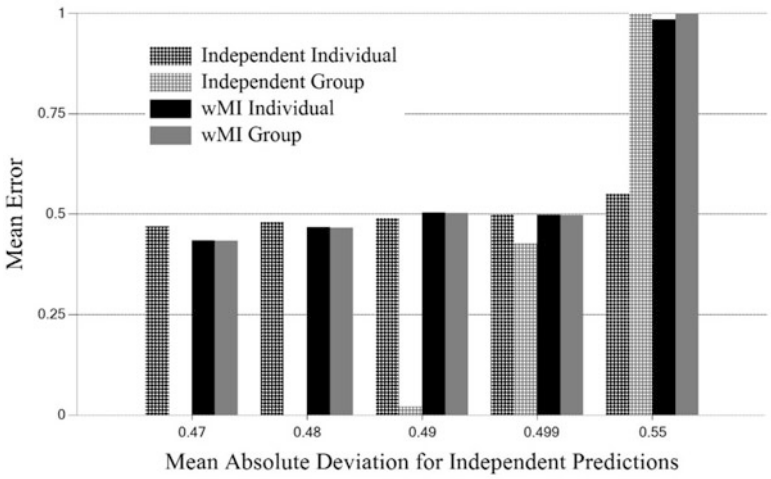


Fig. 1 Binary events, universal access, no experts

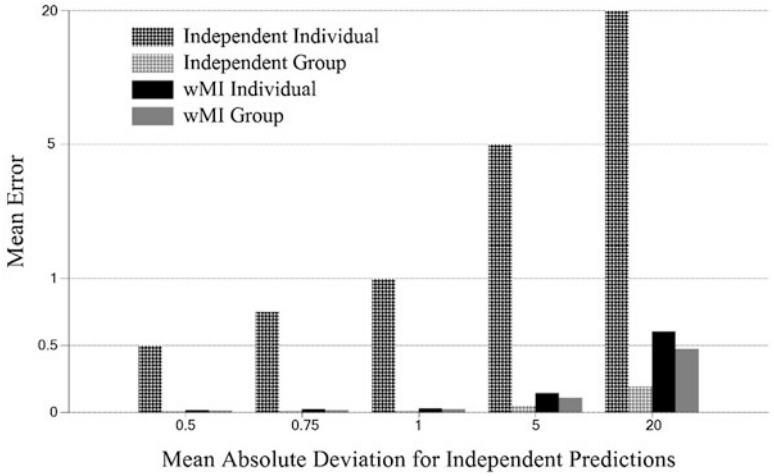


Fig. 2 Real-valued events, universal access, no experts

Noteworthy patterns characterize groups composed wholly of wMIs. In the case of binary prediction games, where the independent unreliability, u, is less than 0.5, we observe no wise crowd effect. In these cases, the populations of wMIs tend to reach an equilibrium state where a single wMI is distinguished as 'most successful', while the remaining wMIs have *identical* success rates, with the result that each prediction made by members of the group is identical to the prediction of the wMI who is most successful. These cases are in contrast to the binary case where u is greater than 0.5, where we observe a 'reverse wise crowd effect', similar to the effect observed in the case of non-imitators (with the additional effect of surging individual error rates). In contrast to the binary case, wMIs exhibit a wise crowd effect in the case of real-valued event sequences. The effect is slightly weaker than in the case of independent predictors, since some diversity is lost through imitation. The reward for the increase in mean group error is a large decrease in mean individual error.

The scenarios represented in Figs. 1 and 2 are unrealistic in assuming that all of the predictors in the group are equally reliable. This unrealistic assumption is biased against wMIs, whose *strength* consists in imitating the predictions of those predictors whose predictions are the most accurate. The simulations represented by Figs. 3 and 4 differ from the ones represented by Figs. 1 and 2, by including a 10 % subpopulation of (expert) independent predictors, with an independent unreliability of 0.1. Note that the independent variable is the independent unreliability of the *non-expert* members of the population, while the mean error rates are derived from the predictions of both experts and non-experts.

Once again, non-imitators have mean individual error rates that converge to their mean independent unreliabilities (which is higher due to the inclusion of subpopulation of highly reliable experts). The impact on the mean individual error

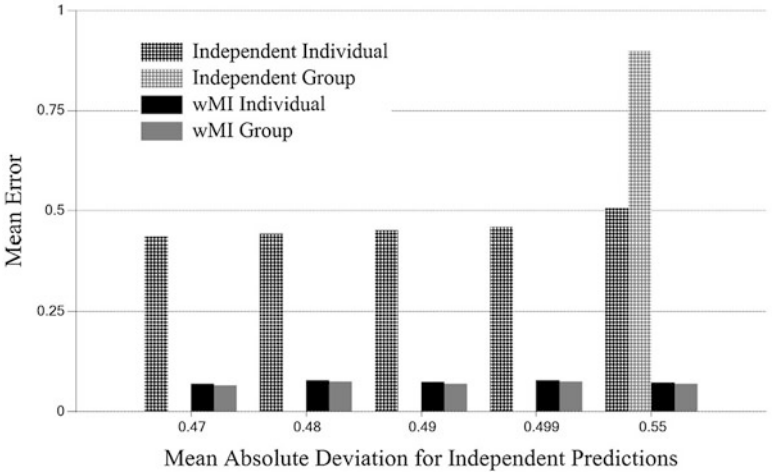


Fig. 3 Binary events, universal access, 10 % experts

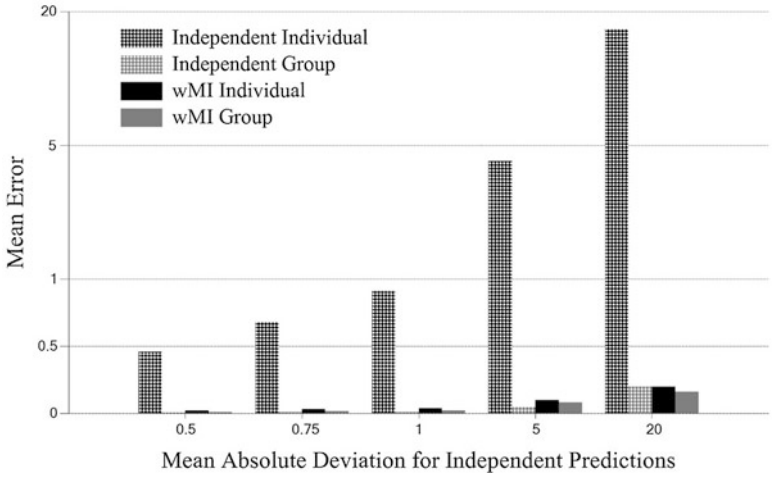


Fig. 4 Real-valued events, universal access, 10 % experts

rate is much greater when we replace the less reliable subgroup of non-imitators with wMIs: The prediction strategy of the wMIs yields the result that the individual error rates of the wMIs approximates the independent unreliability, $u = 0.1$, of members of the highly reliable subgroup (which translates into low group error rates). More generally, applying weighted meta-induction results in lower individual error rates (as compared to independent predictors), so long as we *assume* the wMIs have the opportunity to imitate truth-biased predictors whose independent reliability exceeds their own. The present assumption is typically plausible.

6 Groups with Restricted Accessibility

In the present section, I replicate results from (Thorn and Schurz 2012) in order to illustrate some effects that ensue when members of a group adopt weighted meta-induction, in cases where players only have access to the predictions and success rates of agents in their Moore-neighborhood. In all of the simulations considered in this section, each game lasted 1000 rounds, and had 10 update cycles per round. Within these simulations, the performance of peer imitation is compared with weighted meta-induction. Peer-imitation has some connection to the wise crowd phenomena inasmuch as the predictions of a peer-imitator will be identical to the judgment of the group, in the case where the peer-imitator has access to the judgments of all members of a group. The added effect of peer-imitation over non-imitation, in the case of universal access, is that accurate (or inaccurate) judgments on the part of the group translates into accurate (or inaccurate) judgments on the part of the peer-imitator. In cases where access is limited (as described in Figs. 5, 6, 7, and 8) the connection between the accuracy of the group and the accuracy of the peer-imitator is weakened, but we still observe a tendency of peer-imitators to emulate the judgment of the group, resulting in improved individual accuracy in cases where the group's accuracy is high, and poor individual accuracy where the group's accuracy is poor. Peer-imitation also has a small effect in decreasing the diversity of the group, and thereby on the accuracy of the judgments of the group (in comparison to non-imitation).

Figure 5 presents results for binary prediction games, comparing populations composed wholly of peer imitators to ones composed of wMIs. Figure 6 presents results analogous to Fig. 5 for real-valued predictions.

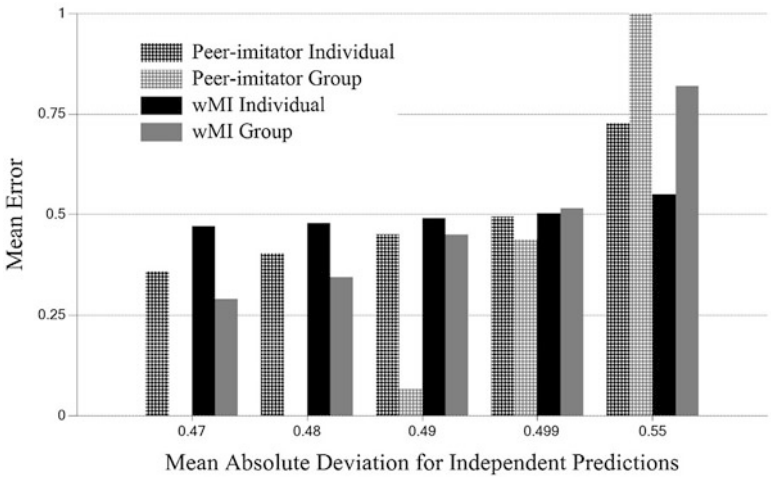


Fig. 5 Binary events, limited access, no experts

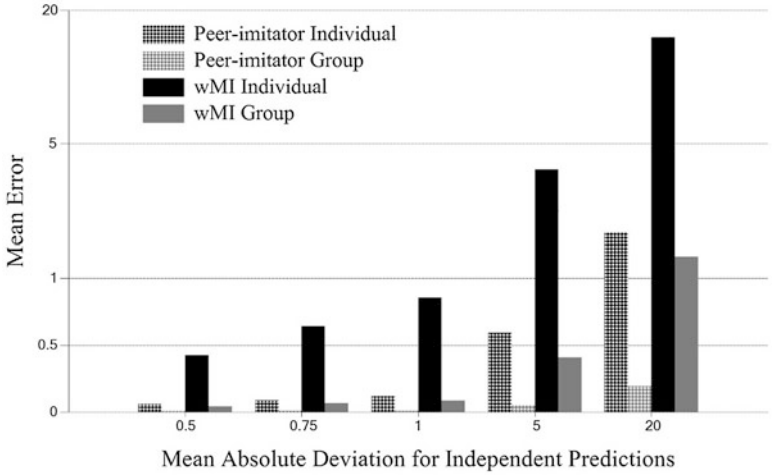


Fig. 6 Real-valued events, limited access, no experts

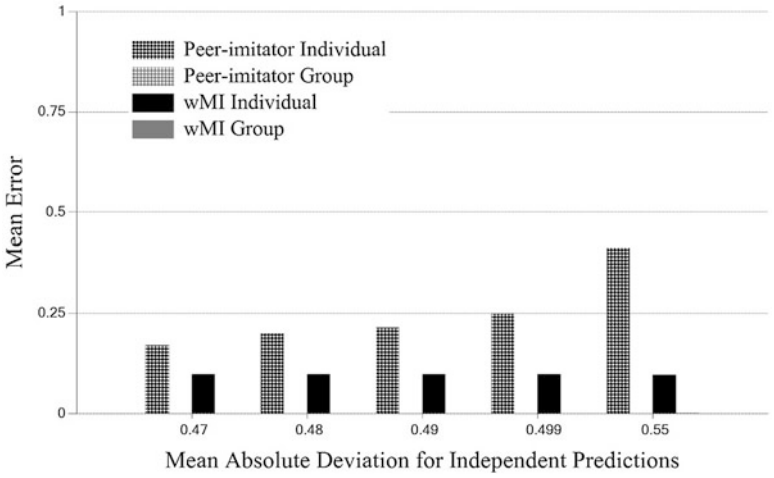


Fig. 7 Binary events, limited access, 10 % experts

In cases where the independent unreliability, u, of all players is identical and truth-biased (i.e., in all the cases described in Figs. 5 and 6, save the case where $u = 0.55$), peer-imitators perform at least as well as wMIs, with respect to individual and group error rates. As it turns out, peer-imitators (within groups of peer-imitators) are incredibly adept in pooling the independent predictions of players with whom they do not have direct access, by taking the average of the predictions neighbors, who took the average of the predictions neighbors, etc. The performance of the wMIs is considerably improved in the case where the wMIs have the opportunity to imitate players with high independent reliabilities. Figures 7 and 8

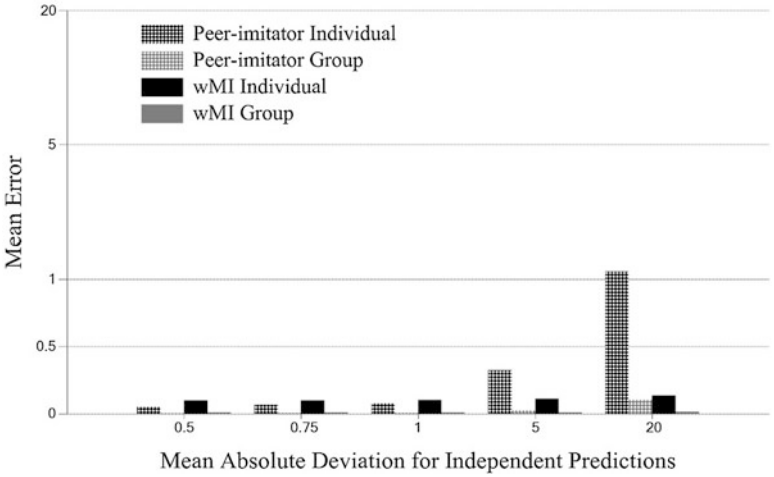


Fig. 8 Real-valued events, limited access, 10 % experts

presents results analogous to Figs. 5 and 6, save that 10 % of the population consists of (expert) independent predictors, with $u = 0.1$.

In all the cases described in Figs. 7 and 8, where we include a subpopulation of expert predictors, we observe that the performance of wMIs matches or exceeds that of peer-imitators.

7 Meta-Induction with Safeguards

While wMI performed well in many of the simulations conducted in (Thorn and Schurz 2012), it performed poorly in others: In binary games with universal accessibility, the mean group error for wMIs was significantly larger than that of independent predictors, in cases where the independent unreliability of the wMIs was low. The mean individual error rates for wMIs were also greater, in the case of binary prediction games where the independent unreliability of the wMIs was high, in the absence of experts. When accessibility was limited, the mean individual and group error rates for wMIs were greater than those of peer-imitators, in all cases where experts were absent and the independent unreliability of the wMIs was low.

I here propose three measures in an attempt to improve the performance of wMIs.[4] The first two measures involve the inclusion of 'virtual' players within the

[4] An alternate variant of weighted meta-induction is considered in (Thorn and Schurz 2012). The variant considered here performs significantly better in several of the situations considered in (Thorn and Schurz 2012). It is also possible to construct situations where the variant from (Thorn

player set available to wMIs as a basis for imitation. First, the MI-players studied in this section, wMI*s, include, as an imitable player, a player that predicts the unweighted average of the predictions of all non-virtual players accessible to the respective wMI*. Second, in the case of binary prediction games, wMI*s consider, as imitable players, the 'inverse' player of each non-virtual players that is accessible to the respective wMI*, where an inverse player always predicts of the opposite of her respective 'non-inverse'. While the former maneuver reflects a self-conscious awareness of the wise crowd phenomena (and attempt to harness it), the latter maneuver aims to make a virtue of systematic error (in the case where $u > 0.5$). The third measure employed by wMI*s is to base their attractivity weights for accessible players on the predictions made in the second to last cycle of each round, which accords with the fact that wMIs are not actually able to imitate the predictions made in the final cycle. The following Figs. 9–12 are analogous to Figs. 1, 2, 5, and 6, and illustrate the effect of the three measures, within those simulations that were the most difficult for regular wMIs, i.e., those situations where the population contained no highly reliable experts. The performance in simulations that include experts is similar to regular weighted meta-induction.

In all of the situations considered here, wMI*s performed better or (almost) as well as independent predictors, and peer-imitators, respectively. Cases where the wMI*s trailed behind their competitors are largely the result of losses earned in the early rounds of the game. While better performance could be achieved within the situations considered here, the required measures would be complicated, and

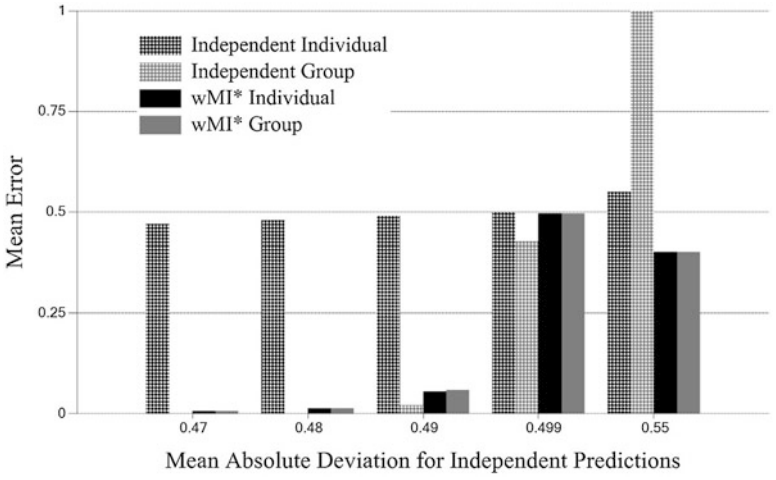


Fig. 9 Binary events, universal access, no experts

and Schurz 2012) performs significantly worse than regular weighted meta-induction, which is not the case for the variant considered here.

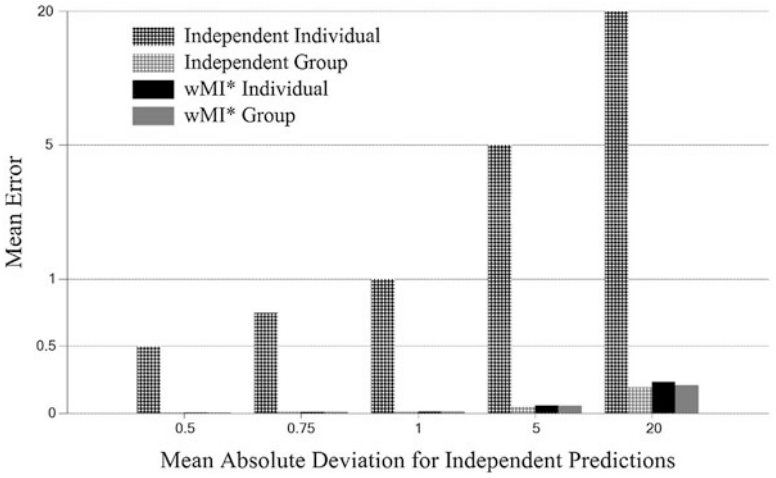


Fig. 10 Real-valued events, universal access, no experts

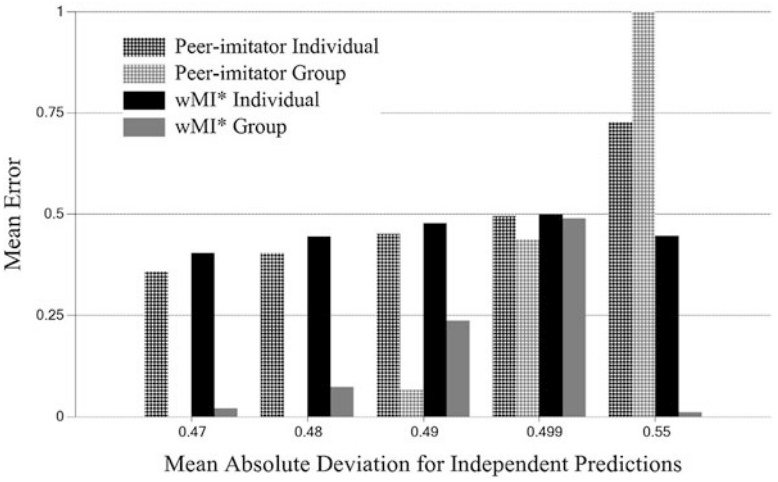


Fig. 11 Binary events, limited access, no experts

would yield only marginal improvements. The performance of wMI*s will also be relatively good within variations of the situations considered here, so long as the (un)reliabilities of the independent predictions of the participating players converge to limits, at a pace commensurate to the length of the respective games.

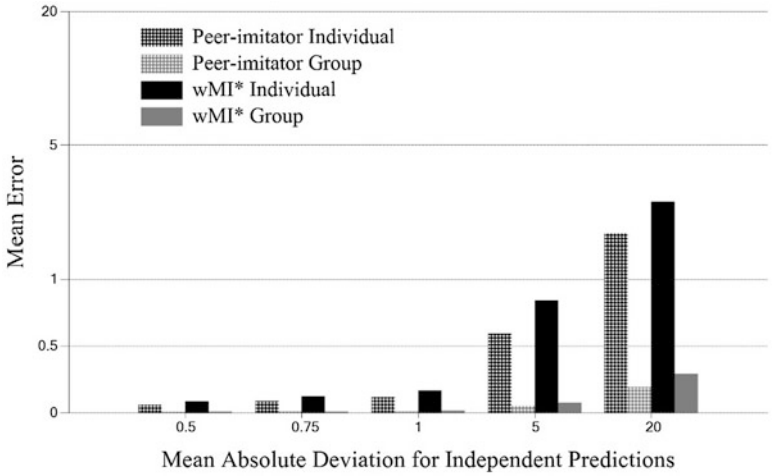


Fig. 12 Real-valued events, limited access, no experts

8 Conclusion

Much recent formal and empirical work on the Wisdom of Crowds has extolled the virtue of *independent* and *diverse* judgment as essential to the maintenance of 'wise crowds'. In contrast, recent work by Schurz (2008, 2009b) demonstrates the optimality of meta-induction as a method for predicting unknown events and quantities. Inasmuch as meta-induction is an imitative prediction method whose application reduces diversity among the predictions of a group, the application of meta-induction may have a negative effect on the accuracy of the average of a crowd's judgment. However, as we saw in the preceding section, it is possible to safeguard meta-inductive methods by simple measures which allow meta-inductive prediction methods to improve their own performance, while simultaneously mitigating their negative impact on group performance.

References

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In G. Tesauro et al. (Eds.), *Advances in neural information processing 7*. Cambridge: MIT Press.

Lorenza, J., et al. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*, 9020–9025.

Page, S. (2007). *The difference – how the power of diversity creates better groups, firms, schools, and societies*. Princeton: Princeton University Press.

Schurz, G. (2008). The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science, 75*, 278–305.
Schurz, G. (2009a). Meta-induction and social epistemology: Computer simulations of prediction games. *Episteme, 6*, 201–220.
Schurz, G. (2009b). Meta-induction. A game-theoretical approach. In C. Glymour et al. (Eds.), *Logic, methodology and philosophy of science* (pp. 241–266). London: College Publications.
Schurz, G. (2012). Meta-induction in epistemic networks and social spread of knowledge. *Episteme, 9*, 151–170.
Surowiecki, J. (2004). *The wisdom of crowds*. New York: Anchor Books.
Thorn, P., & Schurz, G. (2012). Meta-induction and the wisdom of crowds. *Analyse & Kritik, 34*, 339–366.

Is the Equal-Weight View Really Supported by Positive Crowd Effects?

Christian J. Feldbacher

1 Introduction

Adam Elga is one of the most prominent defenders of the so-called equal-weight view in the debate of epistemic peer disagreement, the view that in case of a disagreement between equally well inferentially trained and with the same amount of evidence equipped epistemic agents the difference in opinions should be splitted into equal parts. In fact the equal-weight view presented in Elga (2007) is a little bit more fine-grained, because it also copes with situations of disagreement between epistemic non-peers—may they be no peers due to a lack of equally distributed knowledge about the evidence or due to unequal competences in making adequate inferences (cf. for the more fine-grained version Elga (2007, p. 490)). But for the purpose of our paper it is enough to work with this general characterization: Epistemic peers should meet in the middle.

There are two core-problems of the equal-weight view, namely the problem of spinelessness and the problem of a lack of self-trust (cf. Elga 2007, p. 484). The first problem states that an application of the equal-weight view oughts one to suspend judgement on the issue under discussion too often. The second problem states that an application of this view leads to the implausible consequence "that rationality requires you to give your own consideration of the issue [...] a minor role" (cf. Elga 2007, p. 485). According to Elga the problem of spinelessness is not that pressing, because very often "in real-world cases one tends not to count one's dissenting associates [...] as epistemic peers" (cf. Elga 2007, p. 492). But what of the problem of a lack of self-trust? In Elga's eyes

C.J. Feldbacher (✉)
(Recipient of a DOC Fellowship of the Austrian Academy of Sciences), Department of Philosophy: DCLPS, Universitaetsstr. 1, 40227 Duesseldorf, Germany
e-mail: christian.feldbacher@uni-duesseldorf.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_7

> That problem arose because the equal-weight view entails that one should weigh equally the opinions of those one counts as peers, even if there are many such people. The problem is that it seems wrong that one's independent assessment should be so thoroughly swamped by sheer force of numbers. Shouldn't one's own careful consideration count for more than 1/100th, even if there are 99 people one counts as epistemic peers? (Elga 2007, p. 494)

But

> The short answer is: no. If one really has 99 associates who one counts as peers who have independently assessed a given question, then one's own assessment should be swamped. This is simply an instance of the sort of group reliability effect commonly attributed to Condorcet. [... The equal-weight view] requires one's opinions to be swamped by the majority when one counts a very great many of one's advisors as peers. That is a little odd, but in this case we should follow the Condorcet reasoning where it leads: we should learn to live with the oddness. (cf. Elga 2007, p. 494)

So, his main argument against the self-trust problem seems to be to accept the oddness of a lack of self-trust in order to make profit of a so-called *Condorcet-* or *wise-crowd effect*: If you accept the equal-weight view, you may loose self-trust, but you win a Condorcet- or wise-crowd effect.

In the following sections we will shortly motivate the problem of peer disagreement (Sect. 2) and then characterize the Condorcet- or wise-crowd effects in detail (Sect. 3). Afterwards we will raise two main problems or provisos of Elga's argument (Sect. 4) and end up with a critical conclusion (Sect. 5).

2 The Problem of Peer Disagreement

Classical epistemology is concerned with the notions of 'belief', 'knowledge', 'justification', 'truth', amongst others. These notions are classically explicated with respect to individual agents α_1, α_2 etc. So, e.g., the classical theory of knowledge $\mathscr{K}$ and qualitative belief $\mathscr{B}$ contains some principles like $\mathscr{K}_{\alpha_1}\varphi \rightarrow \varphi$, i.e. what is known is also true, and $\mathscr{K}_{\alpha_1}\varphi \rightarrow \mathscr{B}_{\alpha_1}\varphi$, i.e. what is known by an agent is also believed by the agent etc. These notions are discussed not only qualitatively, but also comparatively and metrically, as, e.g., in Bayesian epistemology, where one introduces the notion of 'degrees of belief' by a subjective probability function p. Well-known problems discussed in this area are, e.g., the problem of how to combine qualitative, comparative and metrical notions via bridge principles, the problem of how to justify rationality constraints on principles for these notions, and the problem of how to deal with multiple degrees of belief of one and the same agent α_1 in the case of belief updating and of different agents α_1 and α_2 in the case of social epistemology in general. The last mentioned problem lead to some new focusing in epistemology, namely to a focusing on the social component of knowledge, by which it is aimed at providing some principles for combining different degrees of belief p_{α_1} and p_{α_2} to one set of degrees of belief $p_{\{\alpha_1,\alpha_2\}}$. As an example you may think on the stock value prediction of two equally competent or successful stock

traders α_1 and α_2 of one and the same company. So, roughly speaking, they share the same empirical data:

- $p_{\alpha_1}(V_{\alpha_T}(x) = V_{\alpha_1}(x)) = 0.8$ (α_1 is quite sure that her prediction of the event x is correct, where $V_{\alpha_T}(x)$ is the true outcome of x and $V_{\alpha_1}(x)$ is the by α_1 estimated outcome of x)
- $p_{\alpha_2}(V_{\alpha_T}(x) = V_{\alpha_2}(x)) = 0.8$ (α_2 is also quite sure that her prediction is correct).

Since the trader's company has to perform an action, there should be some way of combining both degrees of belief, i.e. α_1 and α_2 have to end up with single degrees of belief $p_{\{\alpha_1,\alpha_2\}}(V_{\alpha_T}(x) = V_{\alpha_1}(x))$ and $p_{\{\alpha_1,\alpha_2\}}(V_{\alpha_T}(x) = V_{\alpha_2}(x))$ and should act according to this pooled opinion. Take, e.g., both traders to agree about the statements of the past, i.e. $V_{\alpha_1}(x_{-1}) = V_{\alpha_2}(x_{-1})$, $V_{\alpha_1}(x_{-2}) = V_{\alpha_2}(x_{-2})$ etc. And take trader α_1 to predict that the stock value will fall, so it holds that $V_{\alpha_1}(x_{-1}) > V_{\alpha_1}(x)$. In addition take trader α_2 to think that the stock value will rise, so it holds that $V_{\alpha_2}(x) > V_{\alpha_2}(x_{-1})$. Such a case is a so-called case of *peer disagreement*, since α_1 and α_2 disagree about the true value of the event x, although both are equally competent, i.e. both were equally successful in the past, and both make use of the same empirical data in their predictions (cf. Feldman 2007). According to these predictions, α_1 probably would suggest selling some stocks, whereas α_2 probably would suggest buying some more stocks. The problem of peer disagreement is now exactly the question whether both can be considered to be (equally) justified and if so, how to decide on this basis of conflicting opinions?

As we have seen in the introductory part, the equal-weight view suggests to affirm the question of considering the conflicting positions as (equally) justified since this is just an implication of considering agents as real epistemic peers. What about deciding on basis of the conflicting opinions? Here the equal-weight view stresses the equality of the justifications against an extra-weight view: To extra-weight an opinion in an overall decision making procedure would be adequate only if different opinions were justified to a different degree, but since in the case of peer disagreement the disagreement is amongst epistemic peers, i.e. amongst opinions of equal justification, also extra-weighting an opinion is inadequate in such a case. Furthermore—and this is not only arguing against an opposing view, but directly arguing in favour of the equal-weight view—in performing a difference-splitting strategy one may also make use of a wise-crowd effect. In the following section we will make the assumptions of this argument explicit.

3 Condorcet Juries and Wise Crowds

There are different strategies discussed in the context of peer disagreement. One strategy is to stick to the disagreement, so α_1's and α_2's degrees of belief remain unchanged. This strategy is sometimes called 'no-difference-splitting strategy'. Another strategy is the one under discussion here, namely to equally weight the opponent's degrees of belief and to end up with a mixed belief (cf. for such a

difference-splitting strategy Page (2007, p. 231) and Elga (2007)). If we assume that α_1 and α_2 have degrees of belief as described above (p_{α_1} and p_{α_2}) and if they were absolutely sure that one of them is right, then their pooled degrees of belief $p_{\{\alpha_1,\alpha_2\}}$ would be according to a equally weighting difference-splitting strategy:

$$p_{\{\alpha_1,\alpha_2\}}(V_{\alpha_T}(x) = V_{\alpha_1}(x)) = p_{\{\alpha_1,\alpha_2\}}(V_{\alpha_T}(x) = \\ = V_{\alpha_2}(x)) = \frac{0.8 + 0.2}{2} = 0.5$$

So, the traders in the foregoing example were as unsure whether the stock value will rise or not, as they were unsure whether the stock value will fall or not and so their suggestion for buying, selling or keeping the stocks would probably depend on their disposition of being an optimistic, pessimistic or neutral gambler.

There are very interesting simulations that suggest not following only one strategy in cases of peer disagreement, but, depending on the purposes at hand, to perform different strategies in such a case. Igor Douven, e.g., made some simulations on simple models of peer disagreement in the empirical sciences where the models consist of three components: disagreement among experts, experimental feedback and noisy data. Very generally summarized, his simulations show that in order to track the true value of an experimental outcome with one's predictions, one could follow different strategies for different situations (cf. Douven 2010):

- In case of unnoisy experimental data, performing a no-difference-splitting strategy is, with respect to the purpose of tracking the true value, of equal value as performing a difference-splitting strategy. In such a case experimental feedback does the job, namely to end up with an agreement about the true value relatively close to the true value. Performing a difference-splitting strategy only diminishes the average time needed for predicting the true value (cf. Douven 2010, p. 150).
- In case of noisy experimental data, performing a difference-splitting strategy is more valuable than performing a no-difference-splitting strategy. But in order to diminish the average time needed for predicting the true value, it can be helpful to switch between difference-splitting and no-difference-splitting strategies (cf. Douven 2010, pp. 151 and 154).

Besides such a heuristics for performing different strategies in cases of a peer disagreement, there are also some more general results for justifying the use of a specific strategy. One and perhaps in a broader context also the best known result regarding this matter is the *Condorcet Jury Theorem*. As we have seen in the introductory part, Elga makes use of this theorem in order to argue against the self-trust problem of the equal-weight view. The theorem states that in the situation of an independent and competent jury that was set up for deciding a yes-no-question, it is more probable that the group's majority decision is correct than the decision of an individual member of the jury. And if the jury size tends to infinity, then the majority decision will be correct. The conditions of the situation are in detail as follows (similar results hold also for situations with weakened conditions: cf.

for references on weakening the independence and competence condition: Dietrich (2008); cf. for weakening the duality condition: List and Goodin (2001)):

- Independence condition: The votes of $\alpha_1, \ldots, \alpha_n$ are independent.

$$p(V_{\alpha_i}(x) = V_{\alpha_T}(x) | V_{\alpha_j}(x) = V_{\alpha_T}(x)) = p(V_{\alpha_i}(x) = V_{\alpha_T}(x)) \quad \forall i \leq n,\ \forall j \neq i \leq n \tag{1}$$

- Competence condition: $V_{\alpha_1}, \ldots, V_{\alpha_n}$ are equally competent and at least better than a fair coin.

$$p(V_{\alpha_1}(x) = V_{\alpha_T}(x)) = \cdots = p(V_{\alpha_n}(x) = V_{\alpha_T}(x)) > 0.5 \tag{2}$$

- Duality condition: The vote is about two options.

$$V_{\alpha_T}(x) \in \{0, 1\} \textit{ and } V_{\alpha_i}(x) \in \{0, 1\} \forall i \leq n \tag{3}$$

For such a situation the Condorcet Jury Theorem holds (cf. Dietrich 2008):

Theorem 1. *Provided the conditions of independence, competence and duality, it holds that:*

- *The probability that the majority's vote regarding x is right is greater than the probability that the individuals are right. In a slogan: 'A group is more competent or wise than the average of its members.' Formally put (with n odd):*

$$p(V_{\{\alpha_1,\ldots,\alpha_n\}}(x) = V_{\alpha_T}(x)) > p(V_{\alpha_i}(x) = V_{\alpha_T}(x)) > 0.5\ \forall i \leq n, \text{where}$$
$$V_{\{\alpha_1,\ldots,\alpha_n\}}(x) = V_{\alpha_T}(x) \text{ iff } |\{i :\ i \leq n \text{ and } V_{\alpha_i}(x) = V_{\alpha_T}(x)\}| > \frac{n}{2} \tag{4}$$

- *The probability that the majority's vote regarding x is right approximates to one by approximation of the group size to infinity. In a slogan: 'Infinitely large groups of independent and competent members are absolutely wise.' Formally put:*

$$\lim_{n\to\infty} p(V_{\{\alpha_1,\ldots,\alpha_n\}}(x) = V_{\alpha_T}(x)) = 1.0, \text{where}$$
$$V_{\{\alpha_1,\ldots,\alpha_n\}}(x) = V_{\alpha_T}(x) \text{ iff } |\{i :\ i \leq n \text{ and } V_{\alpha_i}(x) = V_{\alpha_T}(x)\}| > \frac{n}{2} \tag{5}$$

There are many interesting implications of this theorem. It shows, e.g., that under the described circumstances the competence of the group increases with the competence of its members. But note that the theorem does not state, as is sometimes assumed, that the bigger a group of independent and competent voters is the more wise the

group's decision is. As a counterexample for such a claim just take the following situation: Let the independence and competence condition be satisfied for the voters α_1, α_2 and α_3 and let their votes be as follows: $V_{\alpha_T}(x) = V_{\alpha_1}(x) = 1$ whereas $V_{\alpha_2}(x) = V_{\alpha_3}(x) = 0$. Then, expanding a group $\Gamma_1 = \{\alpha_1\}$ by the independent and competent voters α_2 and α_3 to a group $\Gamma_2 = \{\alpha_1, \alpha_2, \alpha_3\}$ does de facto not enhance the majority's vote. On the contrary, whereas Γ_1's majority decision was right, Γ_2's majority decision is wrong.

As the formal description of the theorem shows, the jury's decision on an event x ($V_{\{\alpha_1,\ldots,\alpha_n\}}(x)$) is a function of the jury members' decision-functions on x ($V_{\alpha_1}(x), \ldots, V_{\alpha_n}(x)$). So the jury's decision method as well as all methods within a difference-splitting strategy are meta methods in the sense that they do not operate on the object level, but on the level of methods, whereas the member's methods are object-based. The theorem states that in specific circumstances performing a meta method is better than performing an object-based method only.

To return to our example of the stock market: The theorem suggests that if there is some disagreement about buying some stocks within a group of independent and competent traders, then the traders should perform a difference-splitting method (here: majority voting) to end up with a probably right decision of the question at hand, namely the question of *to buy or not to buy?*

There is another very general result concerning the justification of a difference-splitting strategy (for the following definitions cf. Krogh and Vedelsby (1995) and Feldbacher (2012), sect.3): Take a group's prediction of the value of an event x to be—similar to the majority voting method in the qualitative case of the Condorcet Jury Theorem—the average of the individuals' decisions (cf. Krogh and Vedelsby 1995, p. 232):

$$V_{\{\alpha_1,\ldots,\alpha_n\}}(x) = \frac{\sum_{i=1}^{n} V_{\alpha_i}(x)}{n} \tag{6}$$

Now, if we want to compare the group's prediction with that of the individuals, then we cannot do this directly since the individuals' predictions may be heterogeneous. That this was not the case in the Condorcet Jury Theorem can be seen in the competence condition above. But we can compare the group's prediction indirectly via the error of the prediction: We introduce a measure for the error of a prediction simply by measuring its difference from the true value and square it in order to achieve equal comparability of under- and overestimations (note that squaring is especially with respect to the following results a quite controversially discussed procedure here). First, we introduce a measure for the error of an individual's prediction (cf. Krogh and Vedelsby 1995, p. 232):

$$E_\alpha(x) = (V_{\alpha_T}(x) - V_\alpha(x))^2 \tag{7}$$

Then one can define a measure for the individuals' error just by calculating the average of the error of each individual (cf. Krogh and Vedelsby 1995, p. 232):

$$E_{\varnothing\{\alpha_1,\ldots,\alpha_n\}}(x) = \frac{\sum_{i=1}^{n} E_{\alpha_i}(x)}{n} \tag{8}$$

And similar to the individual's error we measure the error of the group's prediction simply by measuring the difference of the true value and the predicted value (cf. Krogh and Vedelsby 1995, p. 232):

$$E_{\{\alpha_1,\ldots,\alpha_n\}}(x) = (V_{\alpha_T}(x) - V_{\{\alpha_1,\ldots,\alpha_n\}}(x))^2 \tag{9}$$

One only needs to reformulate the equations to see that the following *The Crowd Beats the Average Law* holds:

Theorem 2 (cf. Page (2007, p. 209) and Krogh and Vedelsby (1995, p. 233)).

$$E_{\{\alpha_1,\ldots,\alpha_n\}}(x) \leq E_{\varnothing\{\alpha_1,\ldots,\alpha_n\}}(x) \tag{10}$$

So, it can be shown that in general the error of a prediction of a group is equal to or smaller than the average error of the group's members, which is again a very general positive feature of applying a meta method in predicting the value of an event x. One can observe furthermore that there are two important factors that influence the group's error. Besides the influence on $E_{\{\alpha_1,\ldots,\alpha_n\}}(x)$ by $E_{\varnothing\{\alpha_1,\ldots,\alpha_n\}}(x)$, there is also some influence by the so-called factor of *diversity of the predictions* of the group's members, where the diversity of an individual's prediction is measured by its distance from the average prediction. And the diversity within a whole group is measured by averaging the diversities of the individuals' predictions (cf. Krogh and Vedelsby 1995, p. 232):

$$D_{\{\alpha_1,\ldots,\alpha_n\}}(x) = \frac{\sum_{i=1}^{n}(V_{\alpha_i}(x) - V_{\{\alpha_1,\ldots,\alpha_n\}}(x))^2}{n} \tag{11}$$

With the help of this measure one can show that the diversity within a group also influences the group's error. *The Diversity Prediction Theorem*:

Theorem 3 (cf. Page (2007, p. 208) and Krogh and Vedelsby (1995, p. 232)).

$$E_{\{\alpha_1,\ldots,\alpha_n\}}(x) = E_{\varnothing\{\alpha_1,\ldots,\alpha_n\}}(x) - D_{\{\alpha_1,\ldots,\alpha_n\}}(x) \tag{12}$$

It therefore holds that the lower the average error or the higher the diversity within a group, the lower the error of the group's prediction.

In the discussion about the adequacy of the equal-weight view, both above stated results are put forward in favour of this view in the way we already mentioned in the introductory part. More explicitly put, the argument runs as follows:

1. Performing the equal-weight view is necessary to make use of a Condorcet- or wise-crowd effect. (since performing equal-weighting just equals satisfying the conditions for the Condorcet- and the wise-crowd theorems)
2. One ought to make use of a Condorcet- or wise-crowd effect! (since it's advantageous compared to the average performance)
3. Hence, one ought to perform the equal-weight view. (with 1 and 2)

In the following section we will discuss especially premise 2 of the argument and show that the constraint of making use of a wise-crowd effect is on the one hand quite counterintuitive from a well-performing agent's point of view. And on the other hand we will stress the fact that an agent's making use of a wise-crowd effect diminishes the advantages of such an effect.

4 Two Problems of Condorcet- and Wise-Crowd Arguments in Favour of Equal-Weighting

Both theorems, the Condorcet Jury Theorem and the last observation about a group's error function, have in common that, provided that the predictions within a group are diverse or independent, then the group's prediction outmatches the individuals' average prediction which is to say that the group's competence exceeds the competence of the individuals' average or that the group's ability is higher than the individuals' average. If the individuals' average performance is high enough—which is of course quite vague—such effects are subsumed under the label 'wisdom of the crowd'.

NB: In the case of the Condorcet Jury Theorem the positive impact of diversity is not that easy quantifyable since there appears no diversity factor explicitly in the equations. Nevertheless one can interpret the independence condition of the theorem as a diversity assumption. An interpretation in this line is provided, e.g., in Ladha (1992) by showing that increasing the correlation between the votes decreases the wise-crowd effect. There are also theorems proven with a more fine-grained diversity factor as, e.g., is done in Stone (2015) where diversity is interpreted as different biases of subgroups to different outcomes.

There are many empirical investigations that try to bring some more sophisticated wise-crowd effects in more specific circumstances to the light. Very straight forward is Francis Galton's observation of a wise-crowd effect in estimating, e.g., the

weight of an ox (cf. the description of the example in Thorn and Schurz (2012, pp. 340ff); a very general, but nevertheless good source for wise-crowd examples is Surowiecki (2005)). But there are also much trickier cases of such an effect. Think of collaborative writing platforms on the internet such as, e.g., Wikipedia. One main stream of analysis of Wikipedia is the "question whether the success of Wikipedia results from a *wise-crowd* type of effect in which a large number of people each make a small number of edits, or whether it is driven by a core group of *elite* users who do the lion's share of the work" (Kittur et al. 2007, p. 1). Since it is not necessary to be a part of a user management system to write or edit contributions, it is very tricky to identify the contributors of one as well as contributors of several articles. Nevertheless one can try to identify contributors by similarity relations between IP-address, *changelogs* etc. The analysis of Aniket Kittur et al., e.g., suggests that in the early times of Wikipedia, an elite group did most of the work whereas nowadays the reliability of the articles and the relative completeness of the whole encyclopedia are due to broad collaborative work (the positive performance of group actions becomes apparent here especially if one changes the metrics from counting reliability relative to available information to counting absolutely available information—such a change in the metrics is undertaken, e.g., in Zollman (2015)). However advantageous group performance may be, one always has to take care that, as already noted for the Condorcet Jury Theorem, just increasing the group size, even by competent agents, does not guarantee an improvement of a wise-crowd effect, nor are group decisions in general the best one can do:

> There is this misconception that you can sprinkle crowd wisdom on something and things will turn out for the best. [. . .] That is not true. It is not magic. (Thomas W. Malone, director of the *Center for Collective Intelligence* at the MIT in an interview, cited in Steven Lohr's *The Crowd Is Wise (When It's Focused)* in *The New York Times*, 2009–07–18)

The main point to be considered with respect to Elga's argumentation is that in performing a difference-splitting strategy one should always keep in mind that the positive feature of the strategy is not a magical thing, but only positive compared to the individuals' average predictions. And it can be negative, e.g. in the case of a still very inaccurate prediction of a group, at least from the best individuals' point of view. So the standards for accepting an advantage of wise-crowd effects at the cost of the oddness of a lack of self-trust seems to be quite low and from a well-performing agent's point of view just inacceptable.

Besides this low standards of acceptance the argumentation of Elga raises a second serious problem: Performing the equal-weight view leads naturally to a consensus or, at least to more conformity within a group of epistemic agents. And since an increase of conformity within a group is nothing else than a decrease of diversity within the group, our detailed discussion of Condorcet- and wise-crowd effects in Sect. 3 should make clear that an application of the equal-weight view diminishes the efficiency of wise-crowd effects: Performing the equal-weight view by the agents $\alpha_1, \ldots, \alpha_n$ results in equal estimations and by this the factor of the diversity within these agents' group, i.e. $D_{\{\alpha_1,\ldots,\alpha_n\}}$ (cf. Eq. 11), vanishes. As a consequence (cf. Eq. 12) also the wise-crowd effect vanishes. Of course, in order

to make use of a wise-crowd effect the group's individuals estimations have to be equalized at some point in time. But as the aforementioned simulations of Douven show it is quite dependent of the actual scenario whether such an equalization is truth-apt or not.

Note that our argumentation assumes one crucial assumption of the context in which the equal-weight view is applied. The crucial assumption is that about shared evidence. One could think that, although at some stage of scientific progress all agents of a group update their degrees of belief into a hypothesis (a specific binary event) h onto an equal level, they may still disagree about the evidence (some other specific binary events) e that supports or undermines the hypothesis, since not in every scenario agents are equally informed about and competent in evaluating the evidence e. Making still use of an equal-weight view after updating according to the equal-weight view would be possible in such a scenario. Take, e.g., a standard scenario of Bayesian update via conditionalization, where the priors are—after recognizing a disagreement—levelled up equally whereas the posteriors of the evidence remain different due to different competences in evaluating it:

$$p_{\alpha_1-prior}(V_{\alpha_1}(h) = 1|V_{\alpha_1}(e) = 1) = p_{\alpha_2-prior}(V_{\alpha_2}(h) = 1|V_{\alpha_2}(e) = 1) = 1$$
$$p_{\alpha_1-posterior}(V_{\alpha_1}(e) = 1) = 1,\ p_{\alpha_2-posterior}(V_{\alpha_2}(e) = 1) = 0.5$$

After updating via conditionalization α_1's posterior degree of belief in h equals her prior conditional degree of belief in h given e since α_1 grasped—correctly or not—evidence e. Agent α_2 on the other side behaves quite differently: She didn't grasp e and by this she is not forced to update similar to α_1. So it might hold that:

$$p_{\alpha_1-posterior}(V_{\alpha_1}(h) = 1) \neq p_{\alpha_2-posterior}(V_{\alpha_2}(h) = 1)$$

And by this both could still make use of a wise-crowd effect in case of a disagreement about h just by splitting the difference in their degrees of belief in h. But note that in such a scenario the evidence is not shared and by this it is no case of a *peer* disagreement. So the not fine-grained equal-weight view considered here simply doesn't apply.

To sum up the argumentation about the second problem one may notice that the provided argument in favour of the equal-weight view and against the problem of self-trust defeats to some extend its own basis.

5 Conclusion

We have seen that the argumentation for the equal-weight view as a difference-splitting strategy in cases of epistemic peer disagreements is twofold. On the one hand there is a line of argumentation which stresses problems of the opposing extra-weight view inasmuch as extra-weighting of one or another opinion amongst a group

of epistemic agents is adequate only if the agent's opinions are differently justified, but since in the case of an epistemic peer disagreement the agents are epistemic peers and by this they have equally well justified opinions, performing an extra-weight view is inadequate.

On the other hand there is a line of argumentation which stresses so-called *Condorcet-* and *wise-crowd effects* in favour of the equal-weight view since averaging among the opinions of epistemic peers results in a better performance than the average single performance would be. We have troubled this line of argumentation here by making two quite problematic assumptions/consequences of it explicit, namely first the assumption that the average performance is a key feature of changing one's degrees of belief in case of a peer disagreement. This assumption is from the best performing agents' point of view quite problematic. And second we showed that as a consequence of performing an equally-weighting strategy one diminishes possible advantages of Condorcet- and wise-crowd effects just by simply reducing diversity in a group.

Due to this problems we come to the conclusion that an argument for the equal-weight view with the help of Condorcet- and wise-crowd effects does not succeed.

Acknowledgements For valuable discussion regarding this topic I'd like to thank especially Christoph Leitner, Gerhard Schurz, and Paul Thorn.

References

Dietrich, F. (2008). The premises of Condorcet's Jury theorem are not simultaneously justified. *Episteme, 5*(01), 56–73. ISSN:1750-0117. doi:10.3366/E1742360008000233. http://dx.doi.org/10.1017/S1742360000000927

Douven, I. (2010). Simulating peer disagreements. *Studies in history and philosophy of science part A, 41*(2), 148–157. ISSN:0039-3681. doi:10.1016/j.shpsa.2010.03.010. http://www.sciencedirect.com/science/article/pii/S0039368110000221

Elga, A. (2007). Reflection and disagreement. English. *Noûs, 41*(3), 478–502. ISSN:00294624. http://www.jstor.org/stable/4494542

Feldbacher, C. J. (2012). Meta-induction and the wisdom of crowds. Comment on Paul D. Thorn and Gerhard Schurz. *Analyse und Kritik, 34*(2), 367–382. ISSN:0171-5860.

Feldman, R. (2007). Reasonable religious disagreements. In L. Antony (Ed.), *Philosophers without god. Mediation on Atheism and secular life* (pp. 194–214). Oxford: Oxford University Press.

Kittur, A. et al. (2007). Power of the few Vs. wisdom ofthe crowd: Wikipedia and the rise of the Bourgeoisie. Technival report, Alt.CHI at CHI 2007. alt.CHI at 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007), San Jose. http://www.viktoria.se/altchi/submissions/submission_edchi_1.pdf

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 231–238). Cambridge: MIT.

Ladha, K. K. (1992). The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science, 36*(3), 617–634. ISSN:00925853. http://www.jstor.org/stable/2111584

List, C., & Goodin, R. E. (2001). Epistemic democracy: Generalizing the condorcet jury theorem. *Journal of Political Philosophy, 9*(3), 277–306. ISSN:1467-9760. doi:10.1111/1467–9760.00128.

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton: Princeton University Press.

Stone, P. (2015). Introducing difference into the condorcet jury theorem. *Theory and Decision, 78*(3), 399–409. ISSN:0040-5833. doi:10.1007/s11238-014-9426-3.

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.

Thorn, P., & Schurz, G. (2012). Meta-induction and the wisdom of crowds. *Analyse und Kritik, 34*(2), 339–366. ISSN:0171-5860.

Zollman, K. J. S. (2015). Modeling the social consequences of testimonial norms. *Philosophical Studies, 172*(9), 2371–2383. ISSN:0031-8116. doi:10.1007/s11098-014-0416-7.

Why the Realist-Instrumentalist Debate About Rational Choice Rests on a Mistake

Christine Tiefensee

1 Introduction

Within the social sciences, much controversy exists about which status should be ascribed to the rationality assumption that forms the core of rational choice theories. In one corner of the ring, we find realists who argue that the rationality assumption is an empirical claim which describes real processes that cause individual action. In the other corner, we see instrumentalists who maintain that the rationality assumption amounts to nothing more than an analytically set axiom, an 'as if' hypothesis or useful fiction, which helps in the generation of accurate predictions. In this paper, I approach the realist-instrumentalist debate from a different angle by bringing a distinctly normative interpretation of rationality to the contest. This understanding submits that, contrary to realist or instrumentalist readings, rationality is a normative concept, with ascriptions of rationality being normative judgements that evaluate agents, their actions and intentional states.[1] I will argue that this interpretation is correct: the rationality concept is normative. My main objective, though, will be to show that once the realist-instrumentalist debate is seen to rest on a mistaken interpretation of the rationality assumption, this debate loses its footing. More generally, then, this paper is driven by the conviction that discussions about rationality within the social sciences, metaethics and the philosophy of science can

[1]To avoid misunderstanding right from the outset, this normative account submits that the rationality concept is normative, not that it is ethical or moral.

C. Tiefensee (✉)
Legal Studies and Ethics Department, Frankfurt School of Finance & Management,
Sonnemannstrasse 9-11, Frankfurt 60314, Germany
e-mail: c.tiefensee@fs.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_8

greatly benefit by taking each other's findings more strongly into account. It thus seeks to contribute to an interdisciplinary approach which aims to bring together different strands of enquiry so as to make best use of their respective insights.

I will start by clarifying what is understood by the rationality assumption together with its realist and instrumentalist interpretations. This will be followed by arguments for the normativity of rationality. I will then turn to my main objective, setting out how the normativity of rationality helps overcome the realist-instrumentalist debate. The paper concludes by examining possible objections to the normative account and its implications for positive uses of normative rationality assumptions in the social sciences.

2 The Rationality Assumption

Rational choice theory is concerned with instrumental rationality. As such, its underlying rationality assumption can be stated as follows:

(RA) A rational agent *A* chooses the best means to attain a specific end, given *A*'s beliefs.

More formalised, we find the following definition in the rational choice literature:

(RA*) A rational agent *A* maximises the expected value of a utility function defined on *C* relative to a subjective probability distribution defined on *B*.[2]

Accordingly, (RA)'s concern is threefold (List and Pettit 2011: 24): It deals with the way in which *A*'s attitudes connect with his environment (attitude-to-fact standards), how *A*'s preferences[3] and beliefs relate to one another (attitude-to-attitude standards) and how these propositional attitudes result in action (attitude-to-action standards). With regard to attitude-to-fact relations, (RA) implies that rationality ascriptions are to be based on an agent's subjective perception of the situation, not its objective description. Concerning attitude-to-attitude relations, beliefs and preferences are required to meet certain coherence criteria, such as consistency and transitivity together with constraints on probability distributions. This makes possible consistent belief sets and orderings of preferences. In accordance with modern utility theory, the content of preferences is left open and not limited to specific considerations. Finally, attitude-to-action standards demand that actions maximise expected value, where any such maximisation is again regarded as relative to *A*'s subjective beliefs. It must be noted that this specification of the rationality assumption is one amongst several possible interpretations. I will return to some such alternatives in the penultimate section of this paper.

How do (RA) and (RA*) relate? I will assume here that (RA*) is a formal representation of RA: It offers a mathematical formalisation of our folk psycho-

[2]This is an adaptation of Binmore's (1998: 360–361) definition.

[3]I will use the terms 'preference', 'desire' and 'end' interchangeably here.

logical approach to intentional, purposeful action (Hausman 2012). Acknowledging that (RA*) is RA's mathematical representation is crucial for our understanding of the realist-instrumentalist debate. For, as Lehtinen and Kuorikoski (2007: 123) stress, since utility functions are nothing more than mathematical representations of preference orderings which are, moreover, open to certain mathematical transformations, (RA*) can only ever be read as an 'as if' claim. Consequently, the realist-instrumentalist debate should not be understood as revolving around (RA*): the question is not whether or not agents really employ the mathematical terms of rational choice theory, have utility functions and calculate expected utilities. Rather, the focal point of the realist-instrumentalist debate concerns the status of (RA) and thus the question of whether or not agents really choose in the way that (RA) envisages. Let us turn to this debate next.

3 Realism vs. Instrumentalism

Realists can be distinguished from instrumentalists by their stances on the following three theses:

(1) (RA) is an empirical claim about unobservables.
(2) It is an open question whether or not actual agents are rational.
(3) Actual agents are rational.

Realists endorse all three claims. According to them, the concept of rationality refers to the property of rationality, just as (RA) depicts causal mechanisms that underlie the decision-making process of human beings (1). Realists thus interpret RA empirically: Although RA concerns unobservables—after all, what we can directly observe is agents' behaviour, not their rationality—ascribing rationality to agents amounts to putting forward an empirical description of their psychologies.[4] This enables (RA) to feature not only in predictions, but also in the explanation of phenomena that interest us. Since (RA) is an empirical claim, tests must show whether or not agents really are rational (2). Yet, realists submit that this is a question that we should answer in the positive: (RA) is, even if maybe not outright true, at least a good approximation of the truth (3).[5]

[4] See Popper (1967) for the peculiar position that the rationality assumption is empirical and indeed false, but that we should nevertheless hold on to it, and Lagueux's (1993) attempt to salvage this position. See Hempel (1962) for the claim that the rationality assumption might be abandoned in tests.

[5] Realists' endorsement of (3) is compatible with the thesis that (RA) is a deliberately *false*, idealised empirical description (Mäki 2000). Realists' claim would then be that actual conditions sufficiently approximate these idealised conditions. Since this variation on realism does not affect my argument, I will neglect it here.

Instrumentalists disagree. Contrary to (1), they argue that (RA) should be understood as nothing more than an 'as if' hypothesis[6] or stipulated axiom which we employ in analytic models so as to develop accurate predictions. Since (RA) makes no claim whatsoever about processes that are operative in actual agents' psychologies, seeking to test (RA) is beside the point. Yet, if we leave our analytic models behind and enquire whether or not actual agents do indeed choose as this axiom envisages, instrumentalists agree with (2) that this is an open question. Nonetheless, they reject (3): Either, they declare that actual agents are not rational,[7] or they stay non-committal on whether or not they are. Either way, they hold that (RA) retains its usefulness as a model-theoretic axiom as long as it helps produce accurate predictions. Whether or not actual agents really do choose as postulated by (RA) is none of instrumentalists' business.[8]

4 The Normativity of Rationality

In contrast to both realist and instrumentalist interpretations, I suggest that (RA) should be understood as a normative claim. Holding rationality to be a normative concept is a widespread, albeit not uncontroversial, position in metaethics.[9] Nor is it unknown in the social sciences, where Harsanyi (1976: 90), for instance, declares that "already at a common-sense level, rationality is a *normative* concept: it points to what we *should* do in order to attain a given end or objective."[10] Two features of the rationality concept support this normative interpretation.

Firstly, rationality ascriptions are in a thin sense normative in that they are based on norms or standards: When determining whether or not an individual behaves rationally, we need to apply the rationality standards mentioned above. What endows these standards with thick normativity, though, is that they do not simply divide choices into the categories 'rational' and 'irrational', but that these categorisations carry inherent positive and negative valence: Rational action makes sense, or can be supported by reasons, or has gone right in an important way,

[6] As Lehtinen (2013) emphasises, only certain interpretations of 'as if' signal an instrumentalist outlook.

[7] Friedman (1953) sometimes appears to take up this stance, but compare Mäki (2000).

[8] Rational choice theory can also be understood as being engaged in the explication of the concept of rationality. If understood along these conceptual lines, analytic and normative readings of the rationality assumption need not be mutually exclusive, as the rationality assumption could be understood as the conceptual specification of the normative concept of rationality.

[9] Again, the term 'metaethics' should not mislead readers into thinking that the normative account interprets instrumental rationality as an ethical or moral concept. Strictly speaking, 'metanormativity' would be a more appropriate label than 'metaethics' here.

[10] Compare Gibbard (1990), Southwood (2008), Broome (2007). See also Hands (2011, 2012) and Grüne-Yanoff and Lehtinen (2012) for discussions of normative interpretations of rational choice theory.

whereas irrational behaviour makes no sense, or cannot be supported by reason, or has gone wrong in an important way. We approve of rational action and criticise irrational choices. Rational and irrational behaviour are, therefore, not normatively on a par. Secondly, normative concepts are characterised by their intimate link to action in that, prima facie, we would expect individuals to be motivated to act in accordance with their normative judgements. This is particularly clear within the moral context: If John judges that one ought to help people in need, say, we would expect him to be inclined to help needy people when the situation arises. The rationality concept also possesses this practical import which is paradigmatic of normative concepts. For example, if John intends to travel to Berlin and agrees that it would be most rational for him to catch the train, we would expect him to step on a train, and not take the car, say. Rationality judgements, then, are not only associated with normative pressure to conform to them, but also subject irrational agents to criticism and the demand to correct failures of rationality in future.

Yet, the rationality assumption is not only normative, it is also constitutive of agency. Donald Davidson (1980, 1984, 2004) indefatigably explains why. When ascribing intentional states to others and trying to understand their actions, we are engaged in the business of interpretation: Against the background of our own beliefs, we want to make intelligible each other's actions and the intentional states that lead to these actions. The linchpin of interpretation is, in turn, provided by (RA). In a nutshell, we impute to agent *A* the desire to drink some water, say, if this desire makes intelligible his drinking the glass of water offered to him, provided that he has the belief that the glass contains water. Similarly, the belief that the glass contains water should be attributed to *A* if this belief makes sense of his drinking the water offered to him, provided that he has the desire to drink water. Finally, *A*'s drinking water will count as an action, provided that his wanting to have a drink of water and believing that the glass contains water are *A*'s reasons for his drinking water. More generally, then, when asking which propositional attitudes should be attributed to agents and how to understand their actions, we are guided by the normative question of which intentional states it would be most rational to have and which action would render the agent most intelligible. To adapt Sellars' (1956/1997) words, interpreting agents thus amounts to embedding their actions within the space of reasons.

This implies that judgements of rationality do not evaluate as rational or irrational independently ascertainable propositional attitudes of belief and desire—judgements of rationality are not normative add-ons to otherwise non-normative statements about agents' intentional states. Instead, our conception of propositional attitudes is itself imbued with standards of rationality: It is they which determine which desires and beliefs to impute to an agent by subsuming an agent's behaviour under a consistent, rational set of propositional attitudes. As such, they determine not only what counts as an action and who counts as an agent, but also how to individuate and ascribe beliefs and desires. Standards of rationality, then, are constitutive of intentional states and agency in that an individual who does not satisfy demands of rationality at some minimal level cannot count as an agent in the first place. In Føllesdal's (1982: 312) words: The "assumption that man is rational is . . . inseparable from other hypotheses that we make about man: that he has beliefs

and values, that he acts, etc. We may in a given case be forced to give it up, but then we have to give up these other hypotheses, too."

5 The Mistake of the Realist-Instrumentalist Debate

It is only a small step to see how the insight that (RA) is a normative, necessary precondition for thought and agency affects the realist-instrumentalist debate. Let us remind ourselves that this debate is sparked off by the following theses:

(1) (RA) is an empirical claim about unobservables.
(2) It is an open question whether or not actual agents are rational.
(3) Actual agents are rational.

Realising that ascriptions of rationality are normative shows that both instrumentalists and realists are misguided, either in their respective take on these theses or in their reasons for adopting them. Contrary to (1), (RA) is a necessary, normative precondition for understanding ourselves as intelligible agents who act on the basis of their intentional states. As such, realists are wrong to conceive of (RA) as an empirical claim, just as instrumentalists are wrong to regard it as nothing more than an 'as if' hypothesis. At the same time, instrumentalists are indeed right to point out that attempts to falsify (RA) miss the mark. However, the reason why questions of empirical confirmation and falsification do not apply to (RA) is not grounded in (RA)'s allegedly analytic status, but in its normative nature: Since (RA) is located in the normative and not in the empirical realm, it cannot be empirically confirmed or disconfirmed. Contrary to (2), both realists and instrumentalists are mistaken in thinking that it is a live question whether or not actual agents are indeed rational. Since rationality is constitutive of thought and agency, being an agent amounts to being a rational agent—without rationality, there is no agency. On the normative reading, then, the question of whether or not agents are rational is closed. Finally, with regard to (3), we can see that instrumentalists cannot withdraw to the evasive stance on (RA)'s truth which they intend to take up. In light of the special status of rationality norms as framework principles for agency discourse, there is no logical space for holding that we merely assume that agents act as if they were rational, without declaring that they are rational. Realists, in turn, are right to endorse (3), yet are wrong to support (3) on empirical grounds. The rationality of agents is not an empirical finding, but a precondition of the interpretive process.

To sum up, the realist-instrumentalist debate is fuelled by the assumption that (RA) is a non-normative claim. Arguments to the effect that rationality is a normative concept which is constitutive of thought and agency show that this is a mistake. (RA) is neither open to empirical confirmation and disconfirmation—as realists would maintain—nor allows instrumentalists to take up a non-committal stance on agents' rationality. The question 'Are actual agents really rational or do we only speak as if they were rational?' is, therefore, beside the point.

6 Objections

This argument may be too quick and simple to be fully convincing, so let us look at some possible objections to it. Here, I will limit my discussion to objections that accept the claim that rationality ascriptions are normative whilst questioning the impact that this normativity supposedly has on the realist-instrumentalist debate. I will focus on three worries that I deem the most pertinent. Many of these objections raise further, often thorny lines of enquiry. Pursuing these enquiries in full would go beyond the scope of this paper, so my responses will have to remain schematic at times.

6.1 First Objection

Any plausible theory of rationality must be able to account for cases of irrationality. After all, not only do cases of irrationality abound, we can even identify the processes that cause irrational decisions—just think of framing effects, weakness of will, cognitive biases or wishful thinking, to name but a few. Bearing in mind that the normative account renders questions about agents' rationality closed, it might be concluded that it clearly fails this requirement and should, therefore, be rejected.

This objection rightfully challenges the normative account to show how cases of irrationality can be reconciled with rationality's constitutive role for thought and agency. To conclude that this challenge remains unmet, though, would be premature. For, whilst the normative account does indeed rule out *global* irrationality, it can allow for cases of *localised* irrationality. To elaborate, it might not always be possible to subsume an agent's behaviour under a set of propositional attitudes which fully satisfies the rationality constraint. If so, we are indeed forced to admit that an agent is, in a certain respect, irrational. As long as these lapses of rationality remain sufficiently limited, though, they need not undermine agency status. How pervasive failures of rationality may be without threatening agency is, in turn, a further, normative question—it enquires how much irrationality we can absorb whilst still being able to conceive of an individual as an intelligible agent. Accordingly, the normative account does acknowledge cases of irrationality, but stresses that this irrationality can only ever be local. Global irrationality entails loss of agency status.

6.2 Second Objection

I have mentioned above that the normative account employs a specific interpretation of the rationality assumption, whilst different specifications are also available. One such alternative could be the following:

(RA**) A rational agent *A* maximises his material self-interest relative to objective probability distributions.

Contrary to (RA), (RA**) strictly constrains the content of preferences by focussing on material self-interest only and requires agents to have objectively true beliefs about their situation so as to count as rational. Here, critics may insist that it clearly is a live question whether or not individuals act in line with (RA**): Whether or not they pursue their self-interest and have correct beliefs about probability distributions is, after all, far from being settled. It might be objected, then, that the conclusions of the normative account either miss their target by wrongly focussing on (RA), or must be limited in scope so as to apply to (RA) only.

Bearing in mind (RA)'s prominence and widespread employment in the social sciences, criticising the normative account's focus on (RA) is not convincing. For the same reason, even if its conclusions applied to (RA) only, this would be a significant result. Still, questions about the scope of its conclusions are pressing, so let us look more closely at the status of alternative rationality specifications such as (RA**). Initially, we must note that just like (RA), (RA**) is neither an empirical claim nor an 'as if' hypothesis, but a normative statement: It puts forward a specific interpretation of the value of rationality. Yet, in contrast to (RA), this suggested understanding is substantive in that it evaluates agents' ends by identifying material self-interest as a rational desire and imposes stricter demands on rational belief. As such, (RA**) does not merely demand, like (RA), that an agent's preferences, beliefs and actions stand in certain structural relations to one another, but also requires rational agents to have propositional attitudes with a specific content. Accordingly, if focus on (RA**) were to revive the realist-instrumentalist debate, this would have to be because of these additional demands on the content of agents' propositional attitudes: The question 'Are actual agents really rational?' would now have to be interpreted as the question 'Do actual agents really have these particular desires/beliefs?'. Here, critics are right: The normative account does not close questions about the specific content of agents' desires and beliefs. For, although it entails that if an individual is an agent, his beliefs and desires will adhere to (RA), it does not settle which beliefs and desires these are. As such, it does make sense to ask 'Are actual agents really self-interested, or do we merely assume that they are?'.[11] However, against the background of the normative account, this question now takes on a different hue. For, since the normative account has shown propositional attitudes and instrumental rationality to be indispensable parts of agency, no fundamental doubts about instrumental rationality, unobservables, testability or propositional attitudes can drive this realist-instrumentalist question. Far from doubting the interpretive process, this query must rather be seen as being firmly located within it by enquiring about its outcomes in the form of specific attitude attributions. Consequently, focus on (RA**) does not

[11] Unlike (RA), (RA**) is thus not constitutive of agency: Not pursuing one's material self-interest, say, does not undermine agency status.

reinstate the original realist-instrumentalist contest, but reduces it to a truncated debate about the outcomes of the interpretive process, not its principles. As a result, it has lost much of its bite.

6.3 Third Objection

The final objection consists of three interrelated components. Firstly, the realist-instrumentalist debate seems to be as much about the status of certain assumptions as it is about the aim of science. According to realists, science aims at true descriptions and explanations of the phenomena that interest us, which cannot be based on mere 'as if' hypotheses. For instrumentalists, science aims at accurate predictions on grounds of parsimonious, generalisable assumptions, of which the 'as if' hypotheses of rational choice theory form one example. The normative account, it might thus be objected, does nothing to address this aspect of the realist-instrumentalist debate. Secondly, critics may object that the normative account assumes the social sciences to be interested in the explanation of individual action, whereas their research clearly focuses on the explanation of macro-phenomena. Whilst the rationality of agents may feature heavily when explaining individual action, these critics explain further, it plays no role within explanations of macro-phenomena, where the main explanatory burden is carried by situational and structural characteristics of the agent's environment (cf. Lehtinen and Kuorikoski 2007). This important distinction and its implications are simply neglected by the normative account. Finally, as these previous objections indicate, social science widely employs theories of rationality in positive explanations and predictions. However, by interpreting (RA) as a normative claim, the normative account seems to preclude such positive uses of rational choice theory.

The first component of this objection is partly right and partly wrong. It is wrong in that the normative account does intervene in the controversy about explanation and prediction insofar as it rules out arguments to the effect that rational choice explanations are impossible because (RA) can only ever be an 'as if' hypothesis. Since fundamental doubts about rationality and propositional attitudes are rejected by the normative account, such worries can no longer provide the grounds for rejecting rational choice explanation. It is right in that considerations about the normativity of rationality neither decide whether science should generally aim at explanation or prediction, nor settle whether there might be legitimate uses of 'as if' assumptions in the social sciences. However, since we should not expect normativity to be a magic wand that we can wave so as to solve all questions, and bearing in mind the multifaceted objectives of science, this is just as well.

Secondly, there is indeed a difference between explanations of individual action and those of macro-phenomena. However, this difference is not categorical. To elaborate, take the example of high cost situations which are often quoted in order to illustrate the relevance of structural constraints. In cases such as these, situational restrictions are taken to be so severe that they determine behaviour and thus render

detailed enquiries into individuals' preferences superfluous. As a consequence, the explanatory burden is carried by structural features, not preferences. But even if so, we must realise that whilst high cost restrictions make it easier to ascribe preferences, they do not make preference ascriptions redundant. For, what allows us to narrow down the content of preferences so considerably in high cost situations is that it is easier to determine which preferences it would be rational for agents to have, given their severe situational restrictions. (RA) is, therefore, not only involved in the explanation of individual action, but is also implicitly at work in structural explanations.[12]

Finally, it is true that it is not obvious how normative considerations of rationality can feature in positive social science, so we must ask how the normative and positive functions of rationality can be reconciled. Luckily, though, much work has already been done on the connection between reason and cause. Since a full account of this connection would go far beyond the scope of this paper, let me merely indicate here what would be required in order to achieve a reconciliation of rationality's positive and normative functions. With regard to explanation, we would have to explain how normative accounts of action that appeal to agents' rationality and reasons relate to causal explanations of action. Any such explanation must include, amongst others, considerations about the connection between reason and cause, the non-normative supervenience base of normative rationality judgements and its link with mental causation, the possibility of psycho-physical laws and questions about normative explanations.[13] If prediction is understood as the symmetric counterpart to explanation, the same thoughts apply. With regard to predictions that do not harbour any explanatory ambitions, though, the normativity of rationality does not appear to pose much of an obstacle. Since in this case, we are not concerned with questions as to *why* predictions are correct, but only *that* they are correct, we can assume anything we like. As long as the generated predictions prove to be accurate, neither the status of these assumptions nor their truth matter. Of course, we may find such non-explanatory predictions highly unsatisfying. The reason for this, though, is not found in considerations about the normativity of rationality, but in general questions about the relation between prediction and explanation.

7 Conclusion

In this paper, I have endorsed and argued for the position that the rationality assumption is normative. I have further suggested that the normativity of rationality

[12]It may be argued that there are cases in which (RA) is totally irrelevant because a specific model is 'robust' regarding its behavioural assumptions, in the sense that its results remain the same no matter whether we assume rational, habitual or random behaviour, say (Lehtinen and Kuorikoski 2007: 127). Why this should still count as a rational choice explanation, though, escapes me.

[13]See Davidson (1980), Leach (1977), Spohn (2002).

helps to overcome the realist-instrumentalist debate about rational choice. The question 'Are actual agents really rational or do we only speak as if they were rational?' is, therefore, beside the point. Still, many questions remain unanswered. Most pressingly, we need to understand better the role of normative considerations of rationality in positive social science. It is this question which should attract our attention, not debates about realism and instrumentalism.

References

Binmore, K. (1998). *Game theory and the social contract* (Vol. 2). London: MIT Press.

Broome, J. (2007). Is rationality normative? *Disputatio, II* (special issue), 161–178.

Davidson, D. (1980). Actions, reasons and causes. In *Essays on actions and events* (pp. 3–20). Oxford: Clarendon Press.

Davidson, D. (1984). Radical interpretation. In *Inquiries into truth and interpretation* (pp. 125–139). Oxford: Clarendon Press.

Davidson, D. (2004). *Problems of rationality* (Vol. 4). Oxford: Oxford University Press.

Føllesdal, D. (1982). The status of rationality assumptions in interpretation and in the explanation of action. *Dialectica, 36*, 301–316.

Friedman, M. (1953). The methodology of positive economics. In *Essays in positive economics* (pp. 1–43). Chicago: University of Chicago Press.

Gibbard, A. (1990). *Wise choices, apt feelings*. Oxford: Clarendon.

Grüne-Yanoff, T., & Lehtinen, A. (2012). Philosophy of game theory. In U. Mäki (Ed.), *Philosophy of economics* (pp. 531–576). Oxford: North Holland.

Hands, D. W. (2011). Normative rational choice theory. http://papers.ssrn.com/sol3/papers.cfm?abstractfiid=1738671. Accessed 23 Apr 2013.

Hands, D. W. (2012). The positive/normative dichotomy and economics. In U. Mäki (Ed.), *Philosophy of economics* (pp. 219–240). Oxford: North Holland.

Harsanyi, J. C. (1976). Advances in understanding rational behavior. In *Essays on ethics, social behavior, and scientific explanation* (pp. 89–117). Dordrecht: Reidel.

Hausman, D. (2012). *Preference, value, choice, and welfare*. Cambridge: Cambridge University Press.

Hempel, C. G. (1962). Rational action. *Proceedings and Addresses of the American Philosophical Association, 35*, 5–23.

Lagueux, M. (1993). Popper and the rationality principle. *Philosophy of the Social Sciences, 34*, 468–480.

Leach, J. J. (1977). The dual function of rationality. In R. E. Butts & J. Hintikka (Eds.), *Foundational problems in the special sciences* (pp. 393–421). Dordrecht: D. Reidel Company.

Lehtinen, A. (2013). Three kinds of 'as if' claims. *Journal of Economic Methodology, 20*, 184–205.

Lehtinen, A., & Kuorikoski, J. (2007). Unrealistic assumptions in rational choice theory. *Philosophy of the Social Sciences, 37*, 115–138.

List, C., & Pettit, P. (2011). *Group agency*. Oxford: Oxford University Press.

Mäki, U. (2000). Kinds of assumptions and their truth: shaking an untwisted F-twist. *KYKLOS, 53*, 317–336.

Popper, K. (1967). The rationality principle. Reprinted in D. Miller (Ed., 1985) *Popper selections* (pp. 357–365). Princeton: Princeton University Press.

Sellars, W. (1956/1997). *Empiricism and the philosophy of mind* (R. Brandom, Ed.). Cambridge: Harvard University Press.

Southwood, N. (2008). Vindicating the normativity of rationality. *Ethics, 119*, 9–30.

Spohn, W. (2002). The many facets of the theory of rationality. *Croatian Journal of Philosophy, 6*, 249–264.

Funding Science by Lottery

Shahar Avin

1 Introduction

Contemporary public support of basic scientific research is conducted primarily via allocation by peer review. Under this mechanism, researchers write descriptions of the projects they would like to pursue, and the proposals are ranked by their peers according to their perceived scientific merit. A ranking of the proposals is thus produced, and funding is awarded from the most meritorious downward, until the funds run out.

Recent empirical evaluations of grant peer review have raised concerns about its operation: Graves et al. (2011) find it is not reliable, and Herbert et al. (2013) find it is very costly. The findings have led Graves et al. to suggest a reconsideration of an old proposal by Greenberg (1998), that will allocate a certain portion of the research funds to researchers at random. This paper presents a philosophical motivation for seriously considering the lottery option, by presenting a causal notion of scientific merit and arguing for difficulties involved in estimation of this quantity. The main source of difficulty considered is the dynamic nature of scientific merit, i.e. the possibility of significant changes in the merit of a research project over a short period of time.

S. Avin (✉)
Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge, CB2 3RH, UK
e-mail: sa478@cam.ac.uk

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_9

2 Grant Peer Review

Grant peer review is the dominant contemporary mechanism for allocating public resources to basic scientific projects. Some aspects of the process are strongly conserved across nations and institutions (Dinges 2005; Graves et al. 2011; NIH 2013; NSF 2013).

Grant peer review often offers significant investigator freedom. Project proposals originate from the investigators, not dictated by the funding body or a central organising committee. The extent to which investigators are free to design projects is limited under various guideline constraints, but there are many opportunities for significant levels of freedom.

As proposals originate from the investigators, they must inform the funding body about the contents and merits of their proposed project. This is often done using a detailed written research plan, accompanied by various supporting documents. Funding bodies seek the expert opinion of one or more scientists in evaluating the merit of the proposed projects. While there are guidelines for component categories of evaluation, the decisions are still significantly subjective, not algorithmic or box-ticking.

Usually assessment is sought from more than one source, e.g. from multiple reviewers or from a mix of internal and external reviewers. The different assessments are always combined in some way to form a single judgement per proposal, which is then compared to the judgements of other proposals. There are never enough resources to fund all projects. As such, comparisons of integrated assessments are used to decide which projects will get funded and which will not.

3 Empirical Evidence for Problems with Grant Peer Review

Two recent empirical studies, presented below, look at the level of variability in the grant peer review decisions, and at the cost of running the peer review scheme.

3.1 Measuring the Variability of Peer Review Scores

How can the effectiveness of peer review be measured? One fairly good measure would be to compare the scores of reviewers to the impact of proposed projects (actual in case of funded projects, counterfactual in case of unfunded projects). Such a measurement would give us an estimate of the validity of the merit scores assigned by reviewers. However, the ability to conduct such studies is very limited (Dinges 2005). The key limitations preventing this kind of study are the lack of information

about the impact of projects which were not funded, and the absence of established indicators for measuring the impact of science.

A weaker evaluation of the validity of peer review scores is to check their consistency: to what extent different panel members agree among themselves about the merit of individual projects. The most thorough measurement published to date of the variability of grant peer review scores was conducted by Graves et al. (2011). The authors used the raw peer review scores that individual panel members assigned to 2705 grant proposals submitted to the National Health and Medical Research Council of Australia (NHMRC) in 2009. In the original funding scheme, these scores were given within panels of seven, nine, or eleven reviewers, and the average score of the panel was used to decide whether a project was funded or not, based on its rank relative to other proposals.

In their analysis, the authors resampled from the original scores to generate counterfactual scores. Thus, if the original scores were consistently low or consistently high, resampling will generate a counterfactual average score similar to the original average score. However, if the original scores featured a mix of high and low scores, the resampling will generate counterfactual average scores in a wide range of values. The authors used the counterfactual scores of each project to derive a score interval, or a range of possible scores that the project may have received had the panel composition been different.

The results of the study showed that overall, 61 % of proposals were never funded (score interval was consistently below the funding line), 9 % were always funded (score interval consistently above the funding line), and 29 % were sometimes funded (score interval straddling the funding line).

The authors claim the results show "a high degree of randomness", with "relatively poor reliability in scoring" (p. 3). The authors suggest further research, including investigating the use of a (limited) lottery:

> Another avenue for investigation would be to assess the formal inclusion of randomness. There may be merit in allowing panels to classify grants into three categories: certain funding, certain rejection, or funding based on a random draw for proposals that are difficult to discriminate. (Graves et al. 2011, p. 4)

The above quote suggests a clear link between variability in scores and a (limited) use of a lottery in funding. This link can be made even more suggestive, if we think of the workings of current funding panels *as if* they were an implementation of the system described in the quote. If we black box the workings of the panel, and just look at the inputs and outputs, we see 100 % of the applications coming in, the top 10 % or so coming out as "effectively" funded, the lower half or so being "effectively" rejected, and the middle group being subjected to some semi-random process. Even if we look into the black box, we can see that the process of expert deliberation, when applied to the middle group, bears strong resemblance to the process of a random number generator: it is highly variable and largely unpredictable.

3.2 Measuring the Cost of Grant Peer Review

The cost of the grant peer review system can be broken down into three components: the cost of writing the applications, the cost of evaluating the proposals and deciding on which application to fund, and the administrative costs of the process. According to Graves et al. (2011), in the funding exercise discussed above the largest of these costs was, by far, the cost incurred by the applicants, totalling 85 % of the total cost of the exercise. The authors used full costing of the review process and administration budget, but only a small sample of applicant reports. To complete their data, a more comprehensive survey was conducted amongst the researchers who submitted applications to the NHMRC in March, 2012. The results of this survey, discussed below, are reported in Herbert et al. (2013).

Based on the survey results the authors estimated, with a high degree of confidence, that 550 working years went into writing the proposals for the March 2012 funding round. When monetised based on the researchers' salaries, this is equivalent to 14 % of the funding budget of NHMRC.

The authors also conducted regression analysis on the survey results. Surprisingly, extra time spent on a proposal did not increase its probability of success. Neither did the researcher's salary, which is an indicator of seniority. The researchers' own evaluation of which of their proposals would be funded bore no significant correlation to the actual funding decisions. The only statistically significant effect on probability of success was that resubmitted proposals were significantly less likely to be funded, when compared to new proposals.

The empirical studies discussed above show that despite high costs, the peer review system leaves an epistemic gap between the information provided in the proposals, and the genuine merit of projects, such that high variability exists for a significant middle group. A possible response would be to accept an inherent uncertainty in the process, and cut costs by introducing a less reliable, but cheaper, allocation mechanism, such as a (limited) lottery, especially if some aspects of the current system already operate in a lottery-like manner. The next sections present a reasoned consideration of this alternative.

4 Worries Regarding Random Allocation

There are some immediate objections that can be raised against the proposal to reduce the amount of merit evaluation in peer review and grant room for chance. The central worry is about effectiveness: if we do not rely on evaluation of merit, we would miss out on good research proposals, and will instead end up funding a lot of mediocre science. Challenging this worry will be the main focus of this paper. For completeness, another group of worries regarding the lottery proposal is discussed below, though these worries will not be treated at length.

An expected effect of greater randomness in funding allocation will be a change in the trajectory of research programmes. Under merit evaluation, it is often possible to receive continuous funding for a successful research laboratory, as long as new results are obtained and published and the technology and methods are considered cutting-edge. In contrast, under certain implementations of a lottery mechanism both successful, unsuccessful and novel programmes will have equal chances to win grants, and the relative portion of funds going to continuous funding will be reduced.

A cluster of worries can be associated with discontinuous funding. Continuous funding offers a measure of freedom that can entice highly-skilled individuals despite lower wages compared to other careers. A move to less continuous funding will lose this advantage and may result in less power to attract talent to science.

Many forms of scientific knowledge are gained slowly over time through practice, and are very hard to transfer efficiently to others. With discontinuous funding there is a real threat of losing this gained expertise and accumulated tacit knowledge, which will have a detrimental effect not just on individuals but also on the research environment.

Unexpected scientific discoveries can occur at any point during a funded research project. If a discovery is made close to the end of a funding period, under discontinuous funding there will be less scope to conduct follow-up research, reducing the payoff from such late discoveries.

Much of contemporary research requires significant infrastructure which is tailored to the research project. Such infrastructure needs to be set up, in a costly and time consuming process, any time a new avenue of research is initiated. Continuous research funding offers higher chances of reusing existing infrastructure, and thus offers an efficiency advantage over the costly set-up costs associated with discontinuous funding.

While these are all important worries, they are ultimately technical in nature, and may be solved using appropriate institutional design and practices that will complement the core proposal of funding by lottery. Not so the worry about efficiency, which is deeply associated with the core of the lottery proposal. The empirical evidence surveyed above suggests this might be less of a worry than originally envisioned, but further progress requires a more detailed conceptual analysis of effectiveness in science funding.

5 Scientific Merit

The supposed advantage of grant peer review over random allocation is its ability to make approximately true comparisons between the scientific merit of alternative research projects. An evaluation of peer review's ability to make such comparisons reliably will require a working definition of scientific merit. But what is scientific merit?

A normative definition of scientific merit can be obtained from the initial rationale for public support of research. While the nature of the relationship between a society and its supported scientists may be complex (Geuna et al. 2003), the often cited motivation for public support of science strongly resembles the argument given by Bush (1945). According to Bush, public support of science leads to improvements in health, security, the economy and quality of life. To account for varying social preferences, a more robust definition is given:

Scientific merit (of a research project) is the extent to which the various causal consequences of the project contribute to well-being.

The above definition is deliberately left ambiguous with regards to the exact meaning, or measurement, of well-being. To state the point more formally, merit assignment $M(P, W)$ is a function that takes two parameters, the research project P, and a specific notion of well-being W, and assigns to them a merit score. The merit score of a project, given a certain notion of well-being, can be thought of as how close the consequences of the project will bring us to the specific notion of utopia that emerges from that particular concept of well-being. Thus, for a given notion of well-being, it is possible (in principle) to make comparisons between alternative research projects.

The definition of merit presented above is directly suggestive of some problems that will be involved in its measurement. It is tempting at this point to ditch the proposed definition and opt for a more tractable one. However, the definition as presented above captures a simple but significant notion, that we as a society devote non-negligible resources to scientific research *because we expect science to make our lives better*. Thus, rather than ditch the definition, let us be explicit about the worries of measurement, and follow them through to their consequences.

6 Difficulties in Measuring Scientific Merit

A major difficulty in evaluating scientific merit, as will be argued below, is that merit evaluations are time-dependant. There are two closely related, but distinct, worries involved in this time dependency. The first worry arises from partial and/or fallible knowledge about a target domain. Information about how a certain domain of research may best be explored is unlikely to be available in full until the domain has already been researched; thus, decisions about the most meritorious approach in a certain domain of research contain a substantial element of uncertainty, and future information gained from research may often show past merit evaluations to have been erroneous, despite relying on the best available information at the time.

A second worry is that domains of research are not static. Especially in domains where human and/or technological interventions are sought, such interventions may change significantly the character of the domain, while research is still ongoing. Thus, merit comparisons that rely on the domain being in one state may turn out to be false for the state of the domain at the time when research is being conducted or when the impacts of research are meant to take place.

In both of the above worries, the concern is that the choice to fund project A rather than project B, due to higher assigned merit, would, in hindsight, turn out to have been less effective than if project B was funded. In such a case we would have been misguided, or ignorant, in our assignment of merit, and the worry is that such ignorance may be pervasive. If such ignorance is pervasive, then the relative lack of effectiveness in lottery funding disappears, and with it the strongest objection to the proposal.

Two other difficulties with the evaluation of merit should be mentioned. First, the information about merit is diffuse. The full range of consequences of a research project play out in a wide arena, spatially, temporally, and contextually. Polanyi (1962) addresses this worry with regards to consequences within science, but a full evaluation of relative project merits will depend on knowledge of information diffusion, technological innovation, policy making, and other realms of expertise that may be far removed from scientific practice. It is a challenge facing the evaluators of projects to gather the necessary expertise required to meet this highly heterogeneous demand for information.

Another difficulty originates from the subjective nature of well-being. Like other public servants, science funding bodies are charged with making decisions on behalf of the public, and with a motivation towards the public's best interest. However, the aggregation of public preferences is a notoriously difficult task even in lay matters, let alone in preferences regarding scientific outputs that may require a significant level of tutoring before the preferences can be considered informed (Kitcher 2011).

The above difficulties, regarding diffuse information and subjective evaluation, may turn out to be merely technical. Unlike the worry regarding time-dependant merit, these difficulties are not unique to science, and apply to other matters of public policy. More work is required to ascertain their significance for effective funding allocation, but such work lies outside of the scope of this paper.

Returning to the problem of merit changing over time, the next sections present an evaluation of the extent of the problem, and its consequences. The investigation proceeds in two stages: the next section presents a historical episode featuring multiple occurrences of rapid merit change, based on the account given by Allen (1975); the section following generalises from the historical example by means of a computer simulation.

7 Discovery of DNA: A Historical Example of Rapid Merit Change

Two threads of the story of the discovery of the structure of deoxyribonucleic acid (DNA) can be traced back to the 1860s: one begins with Gregor Mendel's published work on heredity of characteristics in crossbred strains of the common garden pea, the other with the discovery by Friedrich Miescher of nucleic acid, a hitherto unknown substance which is contained in cell nuclei.

The genetic thread of Mendel's work was picked up in 1900, and started a line of experimental work in genetics, which included the discovery that genes are arranged in a linear order on the chromosomes, and that genes were susceptible to mutations. In 1940 the Phage Group was started, with the explicit purpose of solving the mystery of the nature of the gene.

The biochemical thread of Miescher's work was continued, and by the early 1920s it was known that there were two kinds of nucleic acids: ribonucleic acid (RNA), and deoxyribonucleic acid (DNA). By the late 1920s it was known that DNA was located predominantly in the cell nucleus, whereas RNA was located mainly in the cytoplasm. Since the chromosomes were also located in the nucleus, this suggested a greater importance for DNA in the process of heredity. However, the chromosomes are made up of both proteins and DNA, and the consensus opinion was that genes were probably related to proteins, with DNA playing a secondary role. Part of this belief was based on the smaller number of basic components that make up DNA, only four nucleotides, as opposed to the 21 different amino acids that make up proteins. It was believed at the time that the nucleotides repeated in a simple pattern to form DNA.

In 1944, Oswald T. Avery provided the first direct demonstration that DNA was the genetic material. In a transfer of purified DNA from a normal donor bacterium to an abnormal recipient bacterium, the recipient bacterium transformed into the normal state, and descendants of the recipient also inherited the change brought on by the transferred DNA. However, on the background of the known biochemistry detailed above, the reception of Avery's results was very hesitant, and though wildly circulated, it was not accepted into consensus opinion about heredity.

However, in the late 1940s and early 1950s Erwin Chargaff produced experimental evidence that the relative amount of DNA nucleotides differed between species. Chargaff further showed that pairs of nucleotides, adenine (A) and thymine (T) on one hand, cytosine (C) and guanine (G) on the other, appeared in almost identical concentrations, whereas the relative concentrations of AT to CG differed. Given this changed biochemical background, a similar experiment to Avery's was conducted in 1952 at the Phage Group, by Alfred Hershey and Martha Chase. Their experiment showed that when phages infect bacterial cells, it is only the DNA of the phage that actually enters the cell. This further evidence of DNA's role in transmitting genetic information, and the biochemistry that opened room for it to play this role, was sufficient to influence consensus opinion, and focus genetics research on DNA.

The increasing interest in DNA, detailed above, led several groups to attempt to decipher its molecular structure. In 1953 Watson and Crick published the now-famous paper in Nature, in which they describe the double-helix structure of DNA, and suggest its direct role in supporting life by offering a mechanism for replication. Watson and Crick's result had immediate and dramatic effect, and in the following decade was incorporated, through theoretical and experimental work, into what is now known as the central dogma of molecular biology.

8 Modelling Science Funding Under Dynamic Merit Conditions

It might be argued that the historical episode described above is highly unusual in the history of science, involving a unique paradigm-shifting combination of events. To address this criticism, a model has been developed to capture the essence of dynamic merit changes, extrapolating from the example above to less dramatic, but more frequent, occurrences of merit change. The model is a variation of the epistemic landscape as constructed by Weisberg and Muldoon (2009). In Weisberg and Muldoon's model, a community of investigators sets out to explore a particular topic of interest. The various approaches to investigating the topic are represented in a two-dimensional configuration space, with distance between coordinates representing the similarity of the two approaches represented by these coordinates. Each coordinate (approach) is associated a scalar height (significance or merit), representing the value of pursuing that particular approach in investigation of the topic. The community of investigators performs well when the approaches of maximal merit are rapidly found and pursued.

To model dynamic merit, Weisberg and Muldoon's model has been modified to include time-dependant merit. This has been achieved by adding *trigger events*, such that when a particular trigger situation occurs, a change of merit (height) takes place in the epistemic landscape. In the simulation, three such trigger effects have been included:

- Following Strevens (2003), it is observed that little merit is associated with pursuing an approach that has already been successfully pursued in the past. Thus, whenever an approach is successfully pursued by an investigator, the merit of that approach is set to zero for the remainder of the simulation.
- Following Popper (1959), the value of a discovery is positively correlated with the amount of surprise it generates. Thus, when a significant discovery is made (an approach of merit beyond a certain threshold is first explored), nearby approaches lose some of their merit because they would now lead to less surprising results.
- Given the historical example above, it is clear that advances in one area can lead to new avenues of research in another area. In the simulation, this is represented by the appearance of additional merit in a random location on the landscape, following a sufficiently important discovery (when an approach with merit above a certain threshold is first explored).

In order to compare various funding mechanisms, the model has been further modified to include changes in the population of investigators over time. Three mechanisms which have been included are:

- *Best visible*: periodically, new entrants to the field propose to work on approaches at random locations on the landscape. A central funding body only considers those approaches which are sufficiently similar (near) to previously explored

approaches, and selects from them the most meritorious (highest) candidates. Thus, as time progresses and familiarity with the topic increases, a wider set of approaches is considered viable, and merit selection has a wider pool to choose from. This mechanism was designed as a simple representation of grant peer review, where merit-based decisions rely on the past experience of experts.
- *Lottery*: periodically, new entrants to the field propose to work on approaches at random locations on the landscape. A central funding body chooses from them at random, regardless of the merit of their proposed approaches or whether they lie near or far from historical approaches.
- *Triage*: a combination of *best visible* and *lottery*, this mechanism supports half its candidates based on high merit from projects similar to historical approaches, and half its candidates by lottery from approaches which are dissimilar to past approaches. This mechanism was designed as a simple representation of the proposal by Graves et al. (2011) mentioned in Sect. 3.1.

A visualisation of the simulation, including merit dynamics, is shown in Fig. 1 for the *best visible* funding mechanism, in Fig. 2 for the *lottery* mechanism, and in Fig. 3 for the *triage* mechanism.

9 Simulation Results

The simulation was run on landscapes of various sizes, comparing the relative performance of the various funding mechanisms. The measure of success was the total accumulation of merit, i.e. the sum of merit from all pursued approaches throughout the duration of the simulation. The results are shown in Fig. 4 for a landscape of 50×50 approaches, and in Fig. 5 for a landscape of 200×200 approaches.

The results show that on the smaller landscape *triage* and *best visible* strategies outperform *lottery*, suggesting that for niche or restricted areas of research a peer review approach provides an advantage. In comparison, on the larger landscape the *lottery* and *triage* mechanisms outperform *best visible*, suggesting that in very open areas of research, or in situations where multiple topics can combine into one "super-topic" via interdisciplinary links, peer review loses its advantage and a lottery system becomes more appealing.

10 Discussion and Conclusion

The simulation results given above flesh out a reasonable conjecture, that in wide and largely unexplored areas of research, past experience, and expertise that relies on past experience, is only of limited value. Given the drive within science towards exploration of the unknown and revision of the known, both empirically

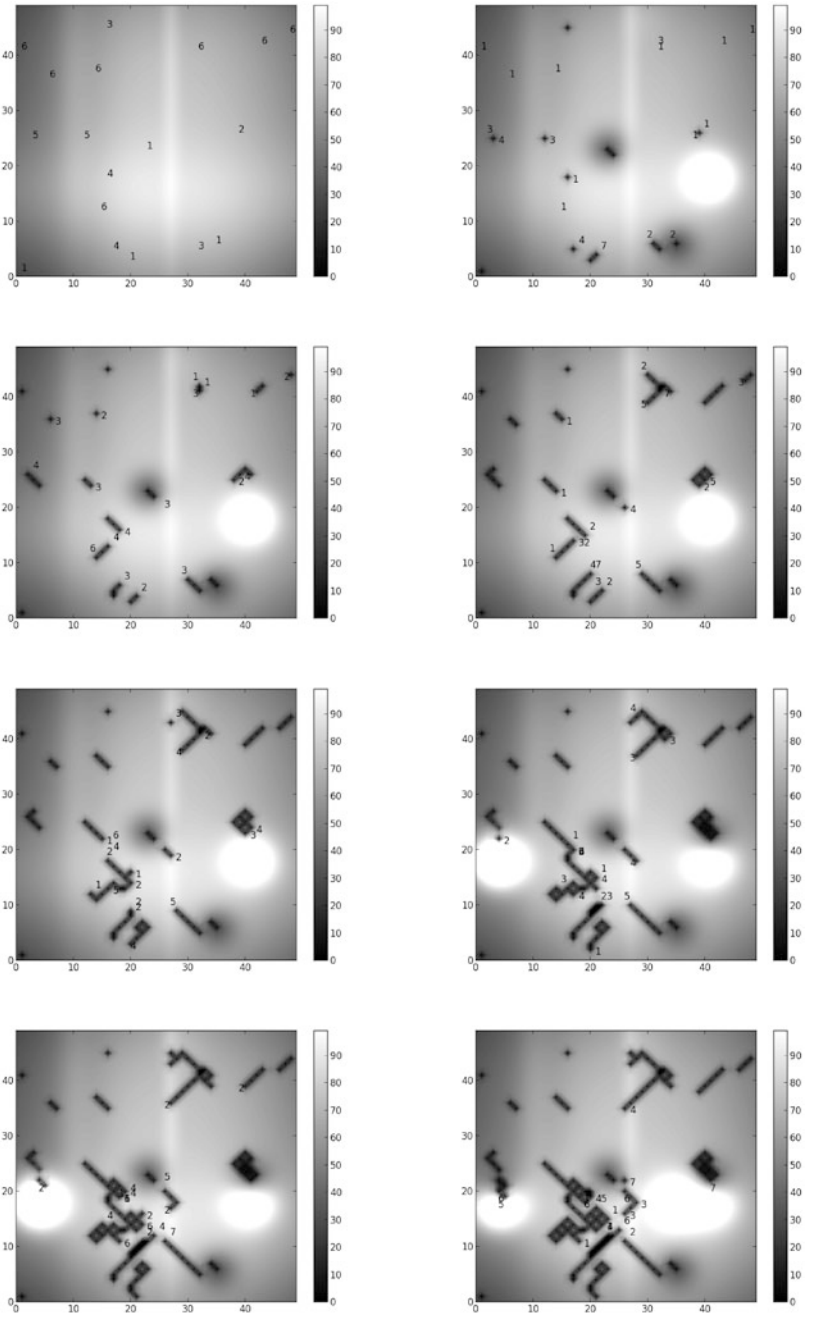


Fig. 1 Simulation of *best visible* funding mechanism on a dynamic landscape. Numbers represent locations of investigators, hue at a coordinate represents its height (brighter is higher)

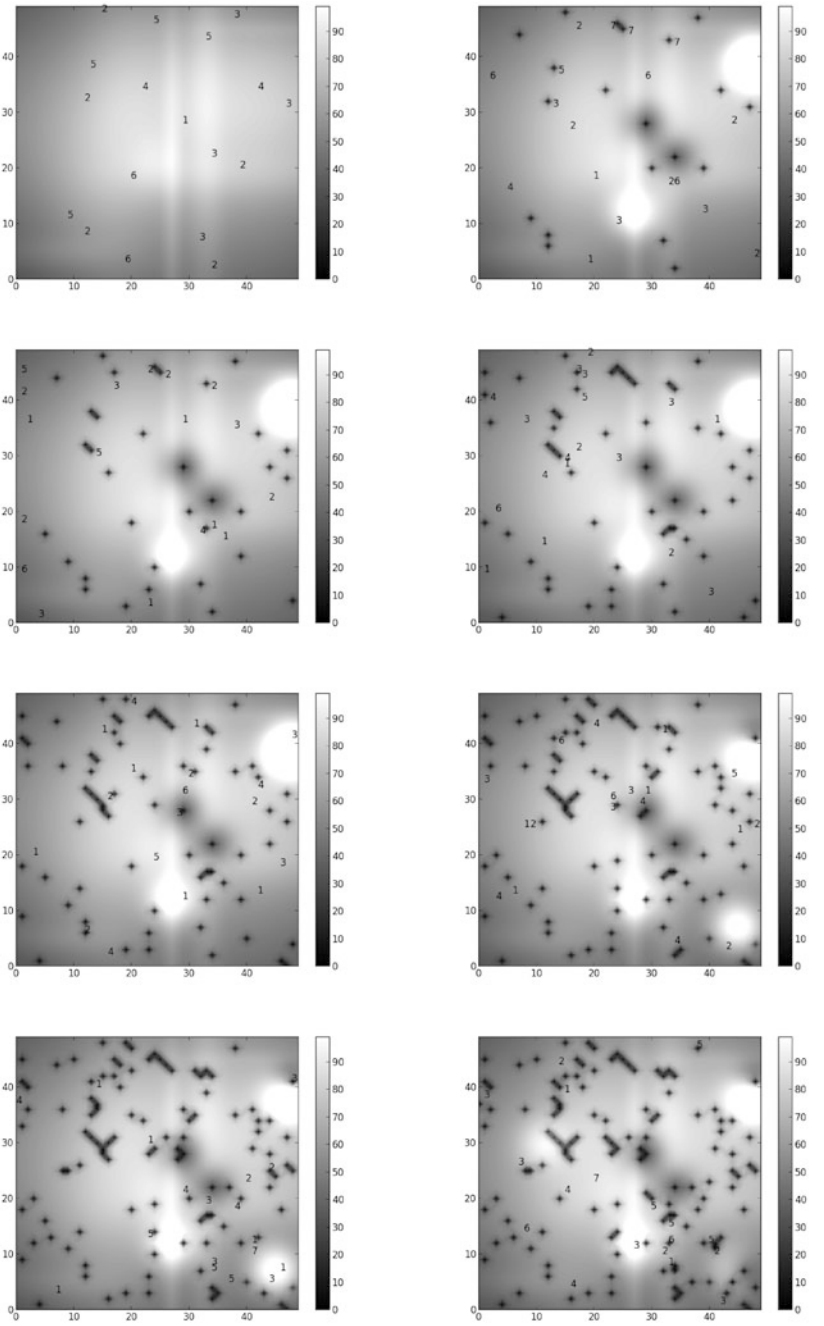


Fig. 2 Simulation of *lottery* funding mechanism on a dynamic landscape

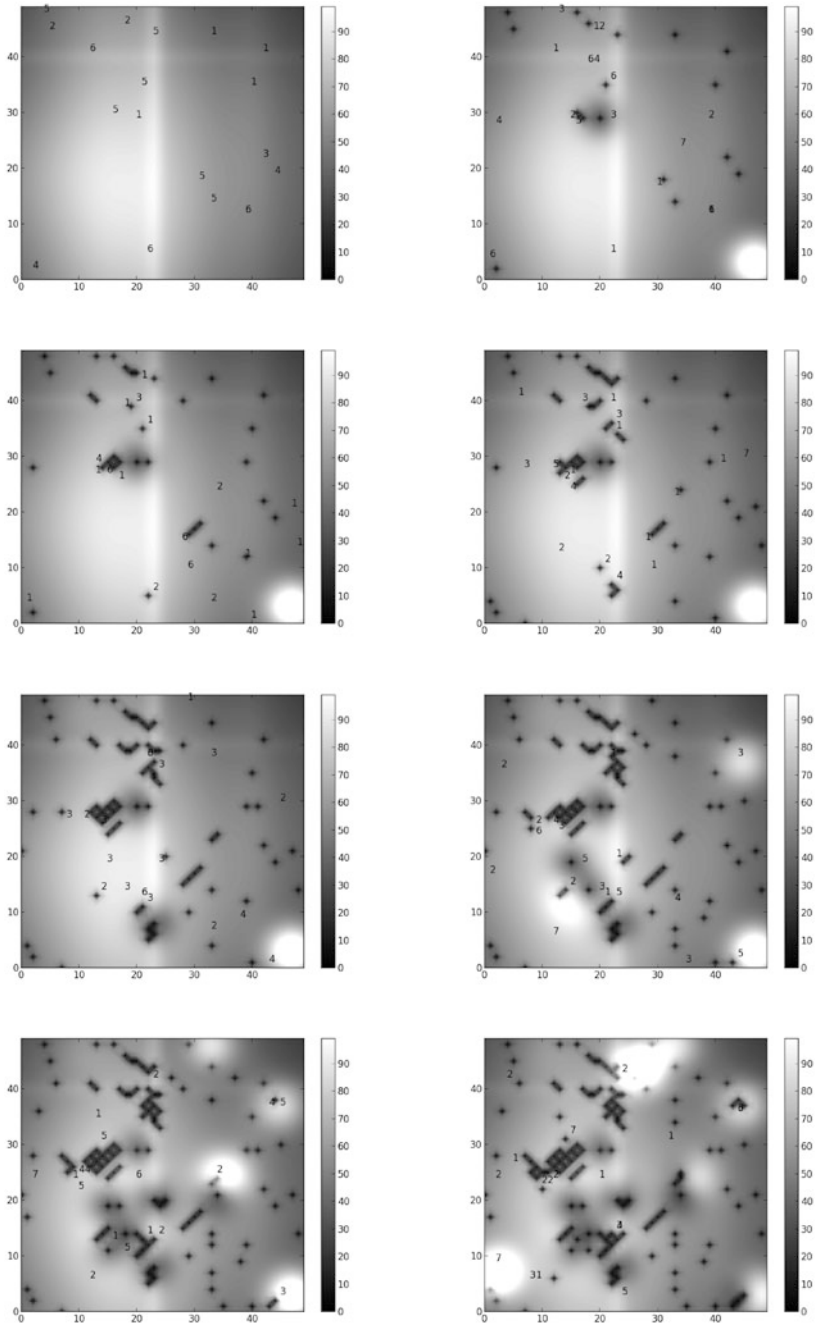


Fig. 3 Simulation of *triage* funding mechanism on a dynamic landscape

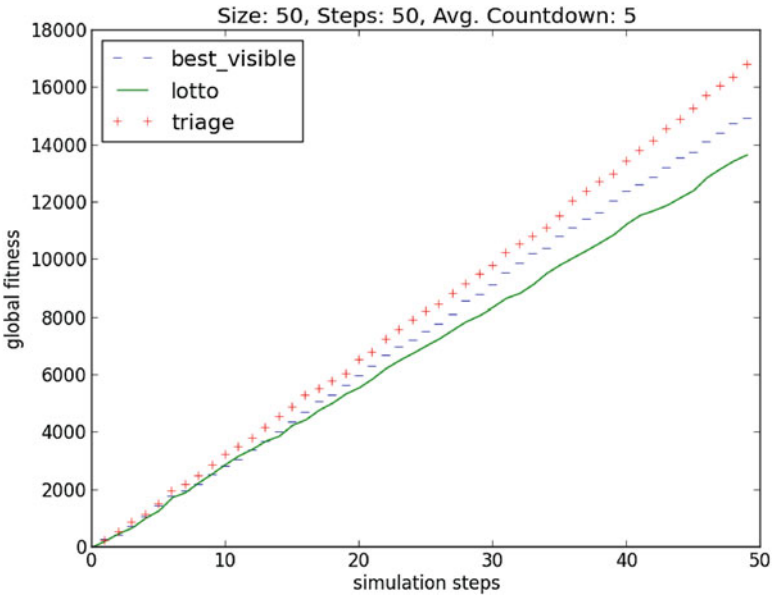


Fig. 4 Comparison of performance for different funding mechanisms over time on a dynamic 50×50 landscape

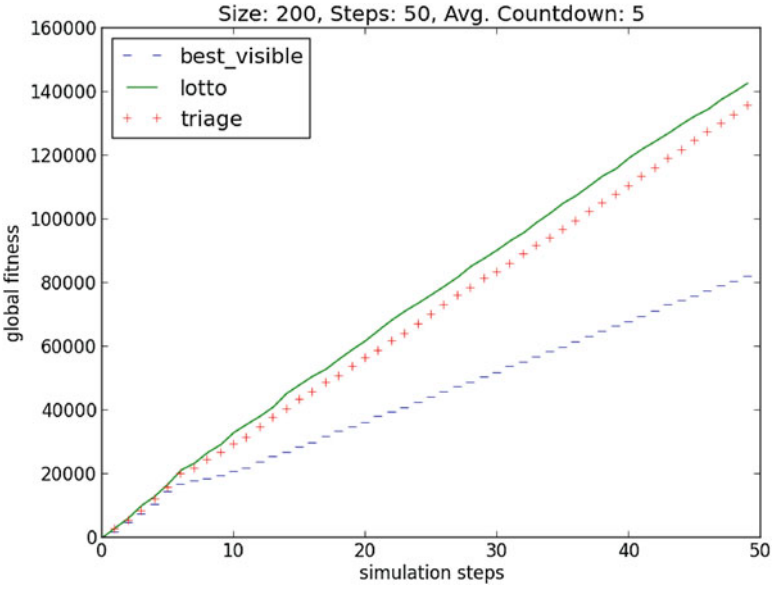


Fig. 5 Comparison of performance for different funding mechanisms over time on a dynamic 200×200 landscape

and theoretically, and the importance of connecting domains of knowledge via interdisciplinary research, it should not come as a surprise that grant peer review is becoming less reliable. The efficiency advantages of random allocation, which at first may seem absurd, are cast in a different light given this explication of reasonable assumptions we already hold regarding the advancement of science.

The relative advantage of the *triage* mechanism on both small and large landscapes suggests a happy medium, as this mechanism combines elements from both peer review and random selection. There could be various ways of implementing such a system in practice, but in all implementations two common features will be present:

– A formal randomisation element will be introduced to select from the pool of proposals amongst those whose merit evaluation is difficult or inconclusive.
– Less information and debate will be required for each proposal, because the exact merit scores of proposals which enter the lottery will no longer matter.

Such a system would reduce the overall cost of the funding exercise, while maintaining overall high effectiveness for scientific research. Rather than worry about lack of reliability in science funding, we should embrace it.

While sketching the core argument for formally including a random element in science funding above, many details of implementation and a discussion of possible consequences have been set aside for lack of space. A more thorough consideration of these matters is presented in Avin (2014), as well as source code for the simulations presented in the previous section.

References

Allen, G. E. (1975). *Life science in the twentieth century* (History of science). New York: Wiley.

Avin, S. (2014). *Breaking the grant cycle: On the rational allocation of public resources to scientific research projects*. PhD thesis, University of Cambridge, Cambridge, https://www.repository.cam.ac.uk/handle/1810/247434

Bush, V. (1945). Science, the endless frontier: A report to the President. U.S. Government printing office, Washington.

Dinges, M. (2005). *The Austrian science fund: Ex post evaluation and performance of FWF funded research projects*. Vienna: Institute of Technology and Regional Policy.

Geuna, A., Salter, A. J., & Steinmueller, W. E. (2003). *Science and innovation: Rethinking the rationales for funding and governance*. Northampton: Edward Elgar Publishing.

Graves, N., Barnett, A. G., & Clarke, P. (2011). Funding grant proposals for scientific research: Retrospective analysis of scores by members of grant review panel. *BMJ, 343*. doi:10.1136/bmj.d4797.

Greenberg, D. S. (1998). Chance and grants. *The Lancet, 351*(9103), 686. doi:10.1016/S0140-6736(05)78485-3.

Herbert, D. L., Barnett, A. G., Clarke, P., et al. (2013). On the time spent preparing grant proposals: An observational study of Australian researchers. *BMJ Open, 3*, e002800. doi:10.1136/bmjopen-2013-002800.

Kitcher, P. (2011). *Science in a democratic society*. Amherst: Prometheus Books.

NIH. (2013). NIH grants policy statement. Accessed Nov 9, 2013, http://grants.nih.gov/grants/policy/nihgps_2013/

NSF. (2013). Grant proposal guide. Accessed Nov 9, 2013, http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg
Polanyi, M. (1962). The republic of science: Its political and economic theory. *Minerva, 1*, 54–73.
Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.
Strevens, M. (2003). The role of the priority rule in science. *The Journal of Philosophy, 100*(2), 55–79.
Weisberg, M., Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science, 76*(2), 225–252. http://www.jstor.org/stable/10.1086/644786

Part III
Values in Science

Researchers Building Nations: Under What Conditions Can Overtly Political Research Be Objective?

Inkeri Koskinen

1 Introduction

In a paper Alison Wylie presented as a keynote address at EPSA 2013, she discusses collaboration with indigenous communities in archaeology, and stresses its epistemic advantages (2015). In this paper I discuss indigenous studies, a young academic discipline which has aims that are similar to those of the kind of collaborative research Wylie discusses, but which goes further in its attempts to bring indigenous knowledge into academia.

Indigenous studies (IS) is an integral part of the international political movement of indigenous peoples that includes groups geographically as distant from each other as the Māoris, the First Nations and the Sámi. The research is overtly political and heavily value-laden: non-epistemic values influence IS in all stages of research. The discipline's proclaimed aims are indigenous self-determination, identity building, mental decolonisation, knowledge building – and even nation building. Today, several colleges and research institutes specialise in IS, including the First Nations University of Canada and Sámi University College in northern Norway. As a part of mental decolonisation and indigenous knowledge building, researchers use

An earlier version of this paper was presented at the International Conference on The Special Role of Science in Liberal Democracy in Copenhagen. I would like to thank Kristina Rolin and the editors of this volume for their valuable comments.

I. Koskinen (✉)
Department of Philosophy, History, Culture and Art Studies, University of Helsinki, Helsinki, Finland

Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences, University of Helsinki, Helsinki, Finland
e-mail: inkeri.koskinen@helsinki.fi

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_10

expressions such as "indigenous knowledge", "indigenous knowledge systems" or "indigenous epistemologies", and claim to be developing "indigenous paradigms" (Porsanger 2010; Stordahl 2008; Smith 1999).

As IS is strongly value-laden, the objectivity of the research conducted would, following the traditional view, be suspect: non-epistemic values should be kept out of the "internal" stages of science (see Douglas 2007)[1]. However, several pluralist philosophers of science have recently questioned the idea that the influence of non-epistemic values on scientific reasoning will necessarily jeopardise the objectivity of research. Helen Longino (2002) maintains that multiple critical points of view are epistemically beneficial and increase the objectivity of a research community, and that this also includes politically motivated views. Heather Douglas (2007, 2009) argues that research that has an impact on human practices cannot and should not be value-free, and that this does not endanger its objectivity. She continues by remarking that several senses of objectivity do not require the value-free ideal – and that they have operationalisable definitions which "can be applied to deciding whether something is actually objective" (Douglas 2007, 131). Value-laden, even overtly political research could thus be found objective in principle. IS represents a critical voice and aspires to influence human practices on many levels, from decision making on ecological issues to the elementary teaching of languages and cultural heritage. It seems to fit the pluralist picture well.

Can research in IS be objective? What is required for it to be objective? Wylie (2012, 2015) has discussed two critiques presented against archaeologists who collaborate with indigenous people. In short, researchers who wish to use indigenous knowledge in their work have been accused of relativism (Boghossian 2006) and of reviving an old, essentialising image of the "native" (Kuper 2003; McGhee 2008). I will claim that these critiques are not based only on the "value-free" meaning of objectivity, but on what Douglas calls *detached objectivity*: "the prohibition against using values *in place of* evidence" (Douglas 2007, 133).

I discuss the same critiques, but from a different angle than Wylie: my focus is not on researchers who collaborate with indigenous peoples, but on IS. In addition, I base my comments on a historical comparison. It is not the first time in European history that researchers have actively taken part in the building of nations. The work conducted in IS resembles in several ways the work that was conducted some 100–150 years ago by ethnologists, linguists, folklorists, historians and others in countries such as Finland that were formed into nation-states in that era. Here I will compare contemporary Sámi IS to Finnish folkloristics over a century ago.

I start with a short description of contemporary Sámi IS and early Finnish folkloristics. Then I continue to the two critiques, which I do not find particularly

[1] According to the traditional view, non-epistemic values may influence research only in the context of discovery, not in the context of justification. However, Douglas (2007, 2009) and Philip Kitcher (2011) argue that researchers need to make some decisions at all stages of research that necessarily include value judgements. For instance, researchers must decide whether the probability for something is high *enough* to warrant a decision, and what is deemed to be *enough* depends partly on context and non-epistemic values.

alarming. Finally, I take into consideration a potential threat to the *interactive objectivity* (Douglas 2007, 135) of IS and discuss the development of young disciplines into objective research communities.

2 Contemporary Sámi Indigenous Studies Compared to Early Finnish Folkloristics

The discipline, or rather the multidisciplinary research programme of IS, is one of the large number of new disciplines born in the last decades of the twentieth century which share the idea that research is always intimately connected to power, is therefore inevitably political, and should be so openly. Thus far indigenous studies has been theoretically heterogeneous, and the theoretical discussion still revolves mostly around the critique of earlier research seen as oppressive.

The Sámi are an indigenous people who traditionally have inhabited Sápmi, the area in Northern Europe also known as Lapland. Sápmi stretches over four countries: Norway, Sweden, Finland and Russia. The political Sámi movement originates from the early twentieth century and was integrated into the growing international indigenous movement in the 1970s. Today, research is an important part of the movement, Sámi IS is represented in several universities, and Sámi University College in Kautokeino, Norway, is wholly dedicated to Sámi IS. Teacher training and linguistics are stressed, as well as the study and revival of the Sámi cultural heritage. Some of the research projects have topics in common with cultural anthropology, folkloristics, comparative religion and ethnology. Other projects include, for example, ecological research related to climate change in the Arctic. Compared to more traditional academic research, indigenous studies aspires to establish a new kind of researcher-subject relationship. Many indigenous researchers conceive themselves as representing not only the academy, but first and foremost their peoples' local ways of thinking, and most of them see themselves as defending their peoples' rights both politically and as researchers. They aim at mental and political decolonisation and at strengthening the self-determination of their people (Seurujärvi-Kari 2011; Stordahl 2008). Jelena Porsanger, the current president of Sámi University College, describes some of the political goals of Sámi IS in the following way:

> On indigenous terms, development is related – among other things – to the strengthening of our societies, the use of our language on different levels, including research and education, the incorporation of our traditional knowledge into resource management in order to secure sustainable use of natural resources, and the reproduction and further development of indigenous knowledge systems transmitted from generation to generation. (Porsanger 2010, 435)

Some Sámi scholars call their work "nation-building" (e.g. Seurujärvi-Kari 2011; Stordahl 2008), thus drawing a parallel between their own work and that of earlier European nation-building projects in which researchers took part. This is not surprising, as the similarities are easy to notice.

Folkloristics is one of the disciplines that contributed in the building of the Finnish nation in the nineteenth and early twentieth century. The nation-building project included the modernisation of the language, the political struggle of adopting it on all levels of education and research, the strengthening of the Finnish national spirit, and the study of Finnish history and traditions. According to the Romantic and Hegelian ideas of the time, a people could not develop into a fully-fledged nation if it did not have a history as a nation. Finland had no such written history, so folklorists found the required history in oral poetry. The high quality of the poetry as literature was also seen as proof of the capabilities of the Finnish people and their ability to progress (Anttonen 2012; Wilson 1976).

3 The Consequences of Relativism

As Wylie (2015) notes, Paul Boghossian has expressed explicit worries about relativism in collaborative archaeology. According to him, a central part of "post-modern" epistemic relativism is "the doctrine of Equal Validity":

> There are many radically different, yet "equally valid" ways of knowing the world, with science being just one of them. (Boghossian 2006, 2)

He continues to claim that this doctrine has "achieved the status of orthodoxy" in many fields, and he has expressed concern about its consequences. As the doctrine leaves no criteria other than ideological ones for adjudicating between radically different ways of knowing, he believes it allows "ideological criteria to displace standards of scholarship" (Boghossian 1996, 14). In other words, he claims that postmodern relativism leads researchers to use values in place of evidence, which goes against the ideal of detached objectivity (Douglas 2007, 133). Finally, he makes a historical comparison to "the recent and sorry history of ideologically motivated conceptions of knowledge – Lysenkoism in Stalin's Soviet Union, for example, or Nazi critiques of 'Jewish science'" (Boghossian 1996, 14), in order to illustrate why postmodern relativism should be resisted.

Wylie sees Boghossian as an "anxious" defender of epistemic foundationalism who "misrecognizes the complexity of actual research practice and ignores the contingency of our evolving epistemic norms" (Wylie 2015). I too believe that he is too hasty in his prediction. Not only does he miss the complex reality of research, as Wylie notes, but I believe that the causal chain he outlines does not hold either. In practice, the doctrine of Equal Validity does not necessarily lead to bad research.

The two short citations Boghossian (2006, 2) gives as examples of the doctrine[2] are ill chosen, as it is questionable if they even represent the doctrine being discussed

[2]Boghossian cites Roger Anyon and Larry Zimmerman, two archaeologists who both collaborate with native American tribes and accept their traditional viewpoints on prehistory as valuable.

(see Koskinen 2011). Boghossian would have found better examples in IS, as one can find an abundance of statements there that do not need to be taken out of context in order to be interpreted as versions of the doctrine. For example, Rauna Kuokkanen (2000) states the following:

> Another objective of an 'Indigenous paradigm' has to do with the recognition and full acceptance of other, alternative epistemologies as being equal to Western systems of knowledge within the academia. As long as indigenous epistemologies are not recognized as being as valuable as Eurocentric epistemologies, Indigenous scholars remain in a marginal, colonial position within the academic institutions. (Kuokkanen 2000, 416–417)

She then illustrates the differences between "Eurocentric" and "Indigenous" epistemologies by referring to "shamanic" knowledge:

> These forms of knowledge which are received through altered stages of mind such as shamanic trances or sudden glimpses of 'seeing', are largely dismissed by Western epistemologies. (*ibid.* 419)

If there is a genuine example of the doctrine Boghossian abhors, this must be it! Nevertheless, and even though Kuokkanen is not alone with her highly questionable ideas, I do not believe that this doctrine will lead to the kind of consequences Boghossian predicts. To illustrate why, I will make a comparison.

Relativistic ideas were also rampant a century and a half ago, when early Finnish folklorists collected oral poetry in order to advance their nationalist ideology. The comparison is not perfect, as the form of relativism that was influential in those days was not the same as the one embraced by some contemporary indigenous scholars. The early nation-builders were influenced by Hegelian notions about *Volksgeist* and by Romantic thought, especially Johann Gottfried von Herder's ideas about language and oral tradition as the expression of a people's spirit. The largely Herderian form of relativism popular amongst the nation-builders of those times was closer to conceptual than epistemic relativism. It was a part of a cluster of ideas that can be described as follows: Nations were taken to resemble living organisms. Each would evolve and grow according to its own nature, and have its own ways of thinking, its own worldview. All thriving nations would take part in the progress of humankind, each in its own unique way. Individuals could achieve their full potential only as part of their own nation, following their nation's spirit. They would never be able fully to understand the thinking of other nations. Language was believed to carry and express the nation's spirit, so individuals had to use their own language if they were to take part in anything creative or progressive. In a foreign language an individual would just poorly mimic foreign ideas. Further, academic research was supposed to be national, to express the national spirit, in order to be actually progressive and not a mere imitation of the achievements of others (Wilson 1976; Abrahams 1993; Anttonen 2012).

If research practices were to follow as straightforwardly from relativistic ideas as Boghossian supposes, this form of relativism would have been likely to lead to nationally self-centered, isolated research programmes that would not have taken part in international discussions. If foreigners were not truly able to understand Finnish research and Finns were not truly able to understand foreign research, there

would not be much reason to pay attention to foreigners. The outcome was, however, surprisingly international. In the late nineteenth and early twentieth century Finnish folklorists developed a historico-geographical method that would be known as "the Finnish method". It led to international recognition of the discipline and became the standard method in folkloristics for decades (Frog 2013).

Influential relativistic ideas do have an impact on research, but not in a vacuum. Rather, they belong – in both of the cases discussed here – to loose clusters of ideas influential at a given time. These clusters can be heterogeneous, even conflicting. For instance, in addition to the aforementioned Herderian and Hegelian ideas, early Finnish folkloristics was influenced by early positivism. In addition, political interests directed the research toward international audiences. Practically all early folklorists were nationalists, yet they disagreed on many issues. The importance of the issues at stake made the theoretical debates heated.

The loose cluster of ideas to which the doctrine of Equal Validity belongs today is no less heterogeneous. For example, certain essentialistic ideas presented in IS are at odds with the doctrine, as they seem to lead to the claim that "indigenous epistemologies" are *not equally valid, but superior to* "Western" science. Even though the doctrine is fairly easy to find in IS, the emerging research practices are not based solely on it (nor are they, I must add, based solely on essentialistic ideas either). It is as yet too early to tell what the outcomes will be, but thus far IS has at least had an influence on the development of collaborative methods in many disciplines and has advanced the study and teaching of indigenous languages. "Shamanic" methods are not in use.

Finally, if we look at the two examples of "ideologically motivated conceptions of knowledge" that Boghossian (1996) mentions, the connection between politically determined research and relativism seems weak. The Nazi critique of Jewish science could perhaps be argued to represent some form of relativism, due to its emphasis on "Aryan" and "Jewish" physics, but Lysenkoists were certainly not relativists – genetics was not called a science, but a "bourgeois pseudoscience". We must conclude that the research practices resulting from relativistic ideas held by researchers are difficult to predict, and that ideologically motivated conceptions of knowledge are not necessarily linked to any form of relativism.

4 Essentialism

IS has been criticised for essentialism. Wylie (2015) quotes Robert McGhee, an archaeologist who claims that Indigenous Archaeology has adopted "a paradigm of Aboriginal essentialism" that he sees as being "derived from the long-discarded concept of Primitive Man" (McGhee 2008, 579). The claim is very similar to one made by Adam Kuper, a social anthropologist, who is an important critic of the concept of "indigenous peoples". He calls it an "ideological makeover of the old idea of 'primitive people'" that is "based on a discredited anthropology" (Kuper 2006, 21). In short, he criticises researchers for using the concept of making a projection,

which has long since been proved to be unfounded by ethnographic research, and for doing so for political reasons (Kuper 2003, 2006). In other words, he claims that researchers who use the concept are not being objective, as they are putting the political rationale first and using the notion against evidence. This would be against the idea of detached objectivity, as researchers are using values in place of evidence.

Essentialist ideas are very apparent in IS. As Gayatri Chakravorty Spivak, who coined the term "strategic essentialism" notes, activist-researchers are prone to resort to essentialism: "my notion just simply became the union ticket for essentialism. As to what is meant by strategy, no one wondered about that" (Danius and Jonsson 1993, 35). Essentialising, often romanticising images are politically effective. The essentialist tendencies that can be observed especially in programmatic statements do not extend only to groups identified as indigenous people. As Arun Agrawal (1995) notes, in IS essentialising generalisations are commonly made about "Western science", which is seen as a uniform ("Cartesian" and "dualistic") knowledge system, and to which indigenous ("holistic") knowledge is presented as an alternative. The indigenous political movement uses the notion of "indigenous people" to unite under one banner people who have had similar experiences *vis-à-vis* colonial and national powers. It seems that the political dichotomy of indigenous people versus "Western hegemony" is reflected in the way in which "Western science" is at times conceived in IS.

Kuper and McGhee refer to the old anthropological notions of "Primitive Man" and "Native". However, the historical comparison they make is not entirely adequate. They concentrate on anthropological notions used in an essentialist way to name the people anthropologists studied – notions that were also suitable for colonial political purposes. The use by contemporary indigenous scholars of the notion of indigenous people, as well as the related, often idealising imagery, is closer to how nationalist ideas and concepts were used in research when building European nation-states. The notion and the imagery used today may resemble those created in early anthropology, but the use to which they are put is closer to the uses which nationalist notions and imagery had in nationalist research some 100–150 years ago. It is illuminating to compare the contemporary nation-building research to research that was equally active in nation-building a century and a half ago; clear differences can be identified.

Early folkloristics, along with other disciplines contributing to building nations in the nineteenth and early twentieth century, was built on essentialist ideas. Finnish folklorists believed themselves to be uncovering the true nature of the nation. Later, especially as a result of the constructionist critique of nationalism presented in the 1980s (Hobsbawm and Ranger 1983; Anderson 1983), folkloristics went through a post-nationalist self-critique comparable to the postcolonialist self-critique in anthropology. The essentialist notions of the early folklorists and other nation-builders in particular have been dismissed, their political projections criticised, and their work seen not as an unveiling but as a building of nations (Abrahams 1993; Anttonen 2012).

Importantly, the constructionist critique of nationalism presented in the 1980s is well known within Sámi IS. In addition, severe criticisms of essentialist ideas and practices have also been made within the Sámi academic community (e.g. Valkonen 2010). In response, some Sámi scholars currently identify their work as taking part in the *building* of a Sámi nation, not as the unveiling of a national essence. The building of "an imagined political community" (Seurujärvi-Kari 2011, 9, see also Stordahl 2008) from varied materials, both traditional and modern, is an openly constructionist idea and not an essentialist one.

5 Dissent

The two critiques just presented are based on the "detached objectivity" sense of objectivity: they claim that IS and similar research in other fields is not objective because the research succumbs to letting values replace evidence. The considerations I have presented to illustrate how these threats can be overcome rest on another sense of objectivity, one that Douglas (2007, 135) names "interactive objectivity". I am optimistic about the reaction of IS to the criticism presented, and believe that in due time it will not be adhering to the kinds of relativism and essentialism described above. My optimism, however, depends on the preparedness of the IS research community to respond to criticism.

Interactive objectivity occurs when a research community reaches an intersubjective agreement on an issue after an intense debate, or when such a community at least follows inclusive procedures that allow effective debates to be had. Longino (2002) has suggested norms or criteria according to which the objectivity of scientific communities can be evaluated, the key point being critical interaction. When discussing collaboration with indigenous communities in archaeology, Wylie (2015) proposes an extension to one of these criteria, namely the norm of tempered equality of intellectual authority, according to which "not only must potentially dissenting voices not be discounted, they must be cultivated" (Longino 2002, 132). Wylie suggests extending this norm outside the confines of scientific or other academic communities:

> In order to counteract the risks of insularity and the effects of dysfunctional group dynamics that can insulate foundational assumptions and norms of justification from critical scrutiny, well functioning [scientific or academic] epistemic communities should actively cultivate collaborations with external communities whose epistemic goals, practices, and beliefs differ from their own in ways that have the potential to mobilize transformative criticism. (Wylie 2015, 207)

IS is often represented as criticism of Western science from the outside. Especially in programmatic texts, the aim of IS is described as twofold. First comes the mental decolonisation and self-determination of indigenous peoples and then, partly in order to succeed at this, comes the critique of Western, ethnocentric research traditions (Smith 1999; Porsanger 2010). As indigenous knowledge systems are stressed in IS, it is often stated that the critique of "Western" research is presented

from the point of view of indigenous communities whose epistemic practices and beliefs differ from those of Western researchers. The role Wylie suggests for indigenous communities in archaeology seems to cohere with the role IS has assumed. However, there is a problem here. Wylie discusses indigenous communities that, being outside academia, can have an asymmetric role *vis-à-vis* researchers: the communities offer criticism that the researchers should take into account, as it can prove epistemically advantageous. IS, however, is not outside academia, and indigenous scholars form research communities. *In order to be objective, these research communities thus cannot adopt only the role of critics.* In accordance with the norms formulated by Longino, *they must also listen to criticism and respond to it*: "Where consensus exists, it must be the result not just of the exercise of political or economic power, or of the exclusion of dissenting perspectives, but a result of critical dialogue in which all relevant perspectives are represented" (Longino 2002, 131).

When trying to undo certain political and economic power assymmetries, IS has shown signs of resorting to the exclusion of certain groups and dissenting perspectives. Vigdis Stordahl summarises accounts of IS debarring non-indigenous researchers:

> Non-indigenous scholars report that they have been asked by indigenous scholars not to conduct research in Sami society. In confidential settings and as "off-the-record" remarks, Sami scholars have reported that they have felt that their research has not been accepted as representing a Sami perspective by the Sami scholars in the Sami research institutions. The meta-communication in operation is that they are not participating in Sami knowledge building. The consequence is that individual scholars find themselves as being almost *persona non grata* and trapped in a classic double-bind situation; whatever you do or say is wrong. (Stordahl 2008, 250)

According to these reports both non-indigenous researchers and indigenous researchers whose perspectives are in some way dissenting, or who perhaps just represent the wrong academic institutions, are meeting with hostile reactions within Sámi IS (Stordahl 2008, 256–259). One culmination point of these attitudes was an influential speech by Nils-Aslak Keskitalo (1974), which has often been interpreted within the Sámi IS community as defining Sámi research as something that only the Sámi could properly conduct: native researchers are always superior to the non-native.

A comparison to early folkloristics shows differences that this time are somewhat worrying: the loose cluster of ideas that inspired early Finnish folklorists included the idea of a *telos* shared by the whole of humankind. Even though individuals were only able to take part in their own national way of thinking, all nations took part in the progress of humanity. The cluster of theoretical ideas influential in IS does not include anything similar. The emphases are rather on power struggles and antagonisms between different perspectives. When these theoretical ideas are connected with experiences of colonialism and earlier anthropological research that resonated with colonial interests, it is no surprise that outsider researchers easily become suspect in IS. However, if such suspicions lead to the exclusion of dissenting perspectives, the Sámi IS community is not objective in the interactive sense of

objectivity. If this were the case, the optimism I presented in the earlier sections would be unfounded. The critique described there is largely presented by outsiders, or when it is presented inside the Sámi community (e.g. Valkonen 2010), it often comes from people who are seen as dissenters within the community.

In two earlier sections I argued that there are reasons for being optimistic with regard to the two critiques presented against IS. My optimism was in both cases based on the belief that the IS research community – or at least the community I have discussed, the Sámi IS community – is reasonably objective in the interactive sense of objectivity: it takes dissenting voices into account. This belief is based on the fact that Sámi scholars have reacted to the critique (as noted above), but also on the observation that the community has opened up in recent years. A growing number of its members protest against exclusive practices. Stordahl's (2008) paper mentioned above is an excellent example of this development.

6 Conclusions

In the course of the last few decades, Sámi IS has established its academic standing. In addition to Sámi University College in Kautokeino, several universities in the Nordic countries have Sámi IS represented in their curricula. Earlier Sámi IS, especially in its programmatic vision statements and certain social practices, has at times been lacking in objectivity, in the sense of both *detached objectivity* and *interactive objectivity*. Both senses proved applicable in my assessment of the discipline. As the discipline has achieved a firmer footing in academia, its interactive objectivity has increased. The research community has been paying more attention to outside criticism and dissenting voices, and has responded to them. As a result, the detached objectivity of the conducted research has also increased. Sámi IS remains as overtly political and value-laden as ever, but is becoming more objective.

The process the Sámi IS community is undergoing is not atypical of overtly political research programmes. In fact, young disciplines in general – or "immature" disciplines in Kuhnian terms – often go through a phase in which they define their agenda, and during which outside criticism is not taken into account as well as it should be according to Longino's norms. When discussing young research programmes or disciplines and emerging research communities, a social, interactive account of objectivity should take these kinds of developments into consideration. Longino's norms are best applied to established academic communities.

In the case of overtly political research programmes, this development is particularly clear, as the difference between theoretical opposition and political opposition is easily blurred, and researchers may interpret all critical comments as expressions of political hostility. Politically engaged research can, however, become objective if the research community opens up to outside criticism and dissenting voices.

References

Abrahams, R. D. (1993). Phantoms of romantic nationalism in folkloristics. *The Journal of American Folklore, 106*(419), 3–37.

Agrawal, A. (1995). Dismantling the divide between indigenous and western knowledge. *Development and Change, 26*(3), 413–439.

Anderson, B. (1983). *Imagined communities: Reflections on the origin and spread of nationalism*. London: Verso.

Anttonen, P. (2012). Oral traditions and the making of the Finnish nation. In T. Baycroft & D. Hopkin (Eds.), *Folklore and nationalism in Europe during the long nineteenth century* (pp. 325–350). Leiden/Boston: Brill.

Boghossian, P. (1996, December 13). What the Sokal hoax ought to teach us. The pernicious consequences and internal contradictions of "postmodernist" relativism. *Times Literary Supplement*, pp. 14–15.

Boghossian, P. (2006). *Fear of knowledge: Against relativism and constructivism*. Oxford: Clarendon.

Danius, S., & Jonsson, S. (1993). An interview with Gayatri Chakravorty Spivak. *Boundary, 2*(20), 24–50.

Douglas, H. (2007). Rejecting the ideal of value-free science. In H. Kincaid, J. Dupré, & A. Wylie (Eds.), *Value-free science? Ideals and illusions* (pp. 120–139). Oxford/New York: Oxford University Press.

Douglas, H. (2009). *Science, policy and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Frog. (2013). Revisiting the historical-geographic method(s). In K. Lukin, E. Frog, & S. Katajamäki (Eds.), *Limited sources, boundless possibilities. Textual scholarship and the challenges of oral and written texts*. A special issue of *RMN Newsletter* (vol. 7, pp. 18–33). Helsinki: University of Helsinki.

Hobsbawm, E., & Ranger, T. (Eds.). (1983). *The invention of tradition*. Cambridge: Cambridge University Press.

Keskitalo, A. I. (1974). Research as an inter-ethnic relation. Paper delivered at the Seventh meeting of Nordic ethnographers, Tromsø Museum in Tromsø, Norway, 29 August 1974.

Kitcher, P. (2011). *Science in a democratic society*. Amherst: Prometheus Books.

Koskinen, I. (2011). Seemingly similar beliefs: A case study on relativistic research practices. *Philosophy of the Social Sciences, 41*(1), 84–110.

Kuokkanen, R. (2000). Towards an "indigenous paradigm" from a Sami perspective. *The Canadian Journal of Native Studies, 20*(2), 411–436.

Kuper, A. (2003). The return of the native. *Current Anthropology, 44*(3), 389–402.

Kuper, A. (2006). The concept of indigeneity – discussion of Alan Barnard's 'Kalahari revisionism, Vienna and the "indigenous peoples" debate'. *Social Anthropology, 14*(1), 21–22.

Longino, H. E. (2002). *The fate of knowledge*. Princeton/Oxford: Princeton University Press.

McGhee, R. (2008). Aboriginalism and the problems of indigenous archaeology. *American Antiquity, 73*(4), 579–597.

Porsanger, J. (2010). Self-determination and indigenous research: Capacity building on our own terms. In V. Tauli-Corpuz, L. Enkiwe-Abayao, R. de Chavez, & J. O. Guillao (Eds.), *Towards an alternative development paradigm: Indigenous peoples' self-determined development* (pp. 433–446). Baguio City: Tebtebba Foundation.

Seurujärvi-Kari, I. (2011). 'We are no longer prepared to be silent': The making of Sámi indigenous identity in an international context. *Suomen Antropologi: Journal of the Finnish Anthropological Society, 35*(4), 5–25.

Smith, L. T. (1999). *Decolonizing methodologies: Research and indigenous peoples*. London: Zed Books.

Stordahl, V. (2008). Nation building through knowledge building: The discourse of Sami higher education and research in Norway. In H. Minde (Ed.), *Indigenous peoples: Self-determination, knowledge, indigeneity* (pp. 249–265). Delft: Eburon.

Valkonen, S. (2010). Essentiaalisten kategorioiden koettelua: Performatiivisuus saamelaisten poliittisen identiteetin ja subjektiuden analyysissa. *Politiikka, 52*(4), 306–320.

Wilson, W. (1976). *Folklore and nationalism in modern Finland*. Bloomington: Indiana University Press.

Wylie, A. (2012). Feminist philosophy of science: Standpoint matters. *Proceedings and Addresses of the American Philosophical Association, 86*(2), 47–76.

Wylie, A. (2015). A plurality of pluralisms: Collaborative practice in archaeology. In F. Padovani, A. Richardson, & J. Y. Tsou (Eds.), Objectivity in science: *New perspectives from science and technology studies* (pp. 189–210). Dordrecht: Springer.

Against the Agnosticism-Argument for Value-Freedom

Anke Bueter

1 Introduction

While most scientists and philosophers of science agree that actual science is not always value-free, many still adhere to the idea that good science is. Proponents of the respective ideal consider value-freedom as a necessary condition for objectivity and thus as decisive for epistemic quality. Opponents of the ideal argue that it is not always possible or preferable to keep science value-free and therefore best to rethink our account of scientific objectivity.[1] Such criticisms of the value-free ideal are often based on some version of underdetermination thesis: Since empirical evidence and further epistemic criteria were insufficient to determine the assessment of scientific theories, hypotheses, models, etc., other factors (such as socio-political or ethical values) would affect this assessment.

Importantly, the respective underdetermination arguments differ in scope and strength. Strong versions hold that all science is underdetermined by all possible evidence, wherefore values necessarily come in. Weaker versions argue that while this is not quite such a global problem, situations of underdetermination do sometimes occur and might in specific cases call for the incorporation of social values into theory assessment. Accordingly, a common defence of the value-free ideal (VFI) is to reject strong versions of the underdetermination thesis in a first step. Second, it is argued that in specific cases where the evidence is in fact insufficient to

[1] Cf., e.g., the conceptions of "strong objectivity" in Harding (1992) or of "social objectivity" in Longino (1990).

A. Bueter (✉)
Institute of Philosophy, Leibniz Universität Hannover, Hanover, Germany
e-mail: anke.bueter@philos.uni-hannover.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_11

decide about the merits of a theory, such decisions should be postponed – not made on the basis of non-epistemic factors. I call this the *Agnostism-Argument* (AA) for value-freedom.

In the following, I will start by characterizing the value-free ideal as well as the agnosticism defence in a bit more detail. I will then argue that the Agnostism-Argument, despite its initial plausibility, is unsuccessful. In particular, I discuss two arguments against it: the problem of a distinction between cognitive and non-cognitive values and the problem of value-laden blind spots in the context of discovery affecting theory choice. In general, what this discussion shows is that not all epistemically relevant decisions can be postponed or determined by further evidence. This defeats not only the Agnostism-Argument but also the value-free ideal itself, since it implies that even a rigorous adherence to the ideal's prescriptions can be insufficient to render theory assessment value-free.

2 Value-Freedom and Agnosticism

A minimal condition of objectivity is that value judgments do not simply override empirical evidence. So far, so good. However, the possibilities for value-influences in science are more complex than this, as is the current VFI and the debate about it. Accordingly, this minimal requirement for objectivity is not at issue here, but will be presumed as necessary. The problem is not one of values *versus* evidence, but starts where the evidence does not unequivocally determine theory assessment. The current VFI specifically aims at such more complicated decisions; it starts from the acknowledgment that empirical data and the internal logical consistency of a theory are necessary but insufficient to guide theory choice. Therefore, it is based on the incorporation of further epistemic criteria, such as external coherence, scope, simplicity, or fruitfulness (cf. the classical list in Kuhn 1977). These criteria are considered as cognitive values – "values", since they do not determine but only guide theory assessment; "cognitive" (or "epistemic"), since they are assumed to promote scientific (cognitive, epistemic) goals and thereby differ from non-cognitive values.

Thus, the current VFI allows for values in science. First, since Max Weber it allows for all sorts of values – cognitive or non-cognitive – affecting decisions on the direction of research or on the use of results. It does not refer to the contexts of discovery or application, but specifies justification, i.e. theory assessment, as its exclusive subject. Here, values are allowed, too – but only cognitive ones.

VFI: Theory assessment should be based exclusively on empirical evidence and cognitive values.

Critics of VFI often argue that evidence and cognitive values do not (always) provide sufficient grounds for theory assessment. The defence via AA then amounts to the following:

AA: In all cases where decisions on the acceptance/rejection of a theory cannot be reached relying exclusively on empirical evidence and cognitive values, such decisions should not (yet) be made.

At first sight, AA seems to be good defence of VFI. Unsurprisingly, several philosophers have argued along such lines. For example, Philipp Kitcher distinguishes three underdetermination theses differing in strength: *Transient underdetermination* refers to situations where the current evidence is insufficient for theory assessment. *Permanent underdetermination* refers to situations in which all possible future evidence will be insufficient, too; *global underdetermination* is the thesis that all theory assessment is permanently underdetermined. Kitcher then rejects global underdetermination and declares permanent underdetermination as unusual. He considers transient underdetermination as more common, but unproblematic for VFI: "Transient underdetermination is familiar and unthreatening" (Kitcher 2001, 30). Although he does not elaborate on this point, it seems clear that it is based on the expectation that these situations will not turn out as permanent ones, giving us the possibility to stay agnostic until the decisive evidence is in.

A more explicit formulation is given by Susan Haack. She first reconstructs the underdetermination argument against VFI and then defends it by AA:

> evidence never *obliges* us to accept this claim rather than that, and we have to accept *something*, so acceptance is always affected by something besides the evidence, which had better be good, progressive values rather than bad, regressive ones. But we *don't* "have to accept something"; if the evidence is inadequate, why not just acknowledge that we don't know? (Haack 1996, 84)

It makes sense to argue that we can simply postpone decisions on the acceptance or rejection of theories; at least as long as there is a reasonable expectation that future research will bring more clarity. Could we thus keep theory assessment value-free by staying agnostic, that is, by admitting we don't know until we in fact know?

In response to AA, it is often argued that there are cases in which we need to act – and therefore to choose on the basis of which theory we are going to act. Thus, it might be impossible or highly undesirable to postpone theory assessment for practical reasons: "withholding judgement is frequently not feasible in real scientific practice" (Kourany 2003b, 24; cf. also Kourany 2003a, b). As Don Howard puts it:

> Debates about long-term underdetermination are irrelevant to the question about science and values because science is done in the here and now, and examples of underdetermination are everywhere to be found in the sciences that make the most difference to human well being. In another 100 years we'll surely know a lot more about how the global climate works. But we have to act now, which means we have to make policy in a setting where the evidence for and against different climate models is not definitive. (Howard 2009, 204)

Kourany and Howard both make an important point about the pressure for action in some areas of transient underdetermination. However, this cannot rebut the essential normative claim of AA, that is, that "[i]f the data fail to agree more with one hypothesis than another, one is *supposed* simply to withhold judgment" (Giere 2003, 20). Proponents of agnosticism could just respond that it is still the best epistemic strategy to postpone judgment, even if it is not simultaneously the best

policy decision to postpone acting.[2] The need for action might trump the need for value-free science in such cases – after all, VFI does not place the value of objective knowledge above all others, but only aims to make such knowledge possible. The pressure of practice does not refute the possibility of saving VFI by agnosticism as a matter of principle; rather, it questions whether this possibility should be given up in certain cases for ethical or political reasons.

Another argument against AA is given by Justin Biddle, who points out that agnosticism does not only come at costs in practical hindsight, but also in epistemic terms. In the tradition of Kuhn and others, he argues that scientists do not actually stay agnostic in the face of transient underdetermination – nor should they. Rather, scientific progress is advanced by commitment in such situations. Only due to such commitments that are actually premature considering the given evidence will the respective theories ever be developed so far as to end transient underdetermination. Agnosticism would thus impede a feature of research that has an important epistemic function.

> The strategy of remaining agnostic in the face of underdetermination does allow for the possibility of epistemically pure science, but it does so at significant cost. (Biddle 2013, 130)

This is a strong argument, and it is actually stronger than Biddle seems to think himself. His argument is that agnosticism comes with the danger of leading to a standstill, albeit a value-free one. If a central mechanism for scientific progress is forestalled by postponing theory choice until sufficient evidence is in, it becomes hard to see in how far the result would still be recognizable as science. In the following, I want to show that this is only one example of a more general point, namely, that it is impossible to postpone all decisions that relate to the evaluation of theories' epistemic merits.

Before spelling this out, let me anticipate a certain objection – that my arguments, as well as Biddle's or Kourany's, trade on certain ambiguities of the notions of

[2] Similar points have been made frequently against the argument from inductive risks, which holds that in cases where errors would have foreseeable consequences if applied in practice, value-judgments on the severity of these consequences affect or ought to affect the assessment of the respective theories (cf. Rudner 1953; Douglas 2000). It is often encountered by a differentiation between belief and action, i.e. the acceptance of a hypothesis and the decision to act upon it (cf. e.g. Jeffrey 1956; Levi 1960; Mitchell 2004; Betz 2013). Defenders of VFI usually argue that scientists should just acknowledge the existing uncertainties (and communicate them to policy-makers). In cases where the need to act prohibits postponing policy-decisions until more evidence is given, such decisions need to be made based on a weighting of the different risks. Thus, things can get messy – but that is a problem of the application of science, not a systematic reason against the epistemic ideal of value-freedom. One way of responding to these arguments is similar to the one I will spell out below: Not only decisions concerning the assessment of developed theories regarding given empirical data are relevant here; instead, this development and therefore the respective theories as well as the generated data are already affected by numerous methodological decisions (e.g., the determination of significance levels) which are underdetermined by the evidence as well (cf. Douglas 2000; Wilholt 2009). In a nutshell, the decisive steps occur *before* we even get to the problem of acceptance or rejection.

acceptance, commitment, theory choice, and the like. The agnosticism-defence is often accompanied by an accusation of conflating two different things: believing in a theory's truth (or warrant) and accepting it as a basis for certain actions. I agree that there is a conceptual distinction to be made here (even though distinguishing two things does not entail their independence). It has recently been proposed to make this distinction along the lines of *belief* in a theory's truth versus its *acceptance* (Elliott and Willmes 2013). I understand the decision to accept a theory as the decision to make it the basis of further work, i.e. to presume it as being true or sufficiently warranted to serve (at least for now) as a stable foundation for further research (cf. also Laudan 1977). While this is similar to the Elliott/Willmes account of *acceptance*, I disagree that the relevant opposite category here is *belief*[3] (among other things because I am sceptical about doxastic voluntarism as well as about the assessibility and thus relevance of such individual mental states). Instead, the distinction at issue here is one between the *reasons for acceptance,* relating to an *epistemic* versus *practical* evaluation of a theory's merits (cf. also Giere 2003; Mitchell 2004). So my argument is not about scientists believing one of two optional theories (or not), but about the decision to make one of them a premise of further research. It is these decisions that are required to be postponed by AA when the epistemic reasons are insufficient and the possibility of their postponing which I want to question now.

3 Refuting Agnosticism

3.1 The Problem of Cognitive Values

It is important to note beforehand that none of the following arguments is based on a strong version of the underdetermination thesis. What they do presume is a hypothetico-deductivist model of scientific testing, which implies that theoretical hypotheses and empirical data do not stand in unequivocal relations. Neither does a hypothesis always logically imply exactly which of its empirical consequences are essential to its evaluation, nor does empirical evidence logically determine whether to accept or reject a theory. If that were the case, there would actually not be any need to supplement the requirement of empirical adequacy by further cognitive values for theory assessment. Thus, VFI itself is based on such a minimal version of underdetermination that comes along with hypothetico-deductivism, since it explicitly invokes such values in order to deal with it.

Now AA argues that in cases where empirical evidence and other cognitive values are still insufficient to assess a theory without reference to other, non-cognitive values, assessment can and should simply be delayed (at least concerning epistemic

[3] AA might then be said to enable value-freedom concerning beliefs, even if acceptance is value-laden.

assessment, see above). The first problem is that this defence reconstructs the problem at hand in a way that takes a central presumption for granted, namely, that there is a legitimate distinction between cognitive and non-cognitive values. This distinction is, however, difficult to establish and highly controversial (for influential criticisms, cf. Longino 1996; Rooney 1992).

This difficulty stems from the general structure of how cognitive values are justified. Characteristics such as simplicity or breadth of scope are not valued for their own sake, but because they are considered to contribute to the goal of science, i.e., to good theories (cf. Laudan 2004). To establish a relation between certain characteristics and this goal, it is thus crucial how this goal is spelled out. One influential line of reasoning here goes back to McMullin (1983), who has coined the term epistemic values and proposed to justify those via their truth-promoting character. Though the assumption of truth (or truthlikeness) being the goal of science is of course a matter of intense debate, we can grant it here for the point of the argument. Even if we assume one superordinate goal of science (truth), it remains questionable how to establish its relation with certain cognitive values. First, there is no analytical or conceptual relation between the notions of truth and simplicity, fruitfulness, etc. Second, this relation can not be settled by empirical testing either. Since the problem of which further criteria to employ as cognitive values only arises on the presumption of different theories which are all largely empirically adequate, any empirical testing would presume a way to compare degrees of truth (proximity) of theories exactly *insofar as they exceed empirical adequacy* – which is precisely what is in question here. To sum this up, cognitive values are justified via their contribution to scientific goals. It is, however, questionable what exactly is *the* goal of science, whether there is one such goal at all (cf., e.g., Kitcher 2001, ch. 6; Elliott and McKaughan 2014), and, if we were to grant these points, how to establish its relation to certain cognitive values.

To relate this to the agnosticism-defence of VFI: If the very criteria for theory assessment cannot be said to be independent from social values, it is of no help to just postpone assessment. AA would have to be complemented by the assumption that further evidence will shed light on what characteristics of theories can count as indicators of epistemic quality. If it were possible to provide empirical support for certain values promoting scientific goals, AA could then be applied to this problem, too.

As the example of truth shows, an empirical support of cognitive values fails because of the required generality of a superordinate goal for all science. If this goal is anything exceeding empirical adequacy, further criteria are needed to indicate goal-conduciveness of theories. In order to empirically test whether certain characteristics of theories are goal-conducive, a way to determine and compare proximity to the goal is needed. This comparison of goal-proximity, however, requires further criteria besides empirical adequacy – the very criteria we are trying to establish.

One way to break out of this circle and enable an empirical evaluation of supposedly cognitive values would be to proceed from more specific research goals. The more concrete these goals are, the easier it will be to determine degrees

of proximity to them. However, it is disputable whether such specific goals will always be promoted by the same values and whether these goals can and should themselves be purely cognitive. The specific research goals must obviously include the aim to generate knowledge (and not only political desirability, for example). Yet, to determine what kind of knowledge about what phenomena is most important might very well be interwoven with socio-political aspects. For example, Longino prominently formulated an alternative list to the Kuhnian one with values that are supposed to promote research that pursues feminist goals, such as making gender visible as a variable or to understand mechanisms of oppression (cf. Longino 1996). The values that are justified via their contribution to such specific goals – which are both cognitive and political – might then be empirically supported, but that does not make them purely cognitive. Instead, we could have, e.g., feminist cognitive values in a specific research context.

To sum up, VFI presumes a problematic distinction between cognitive and non-cognitive values. Staying agnostic would only be helpful if there were a reasonable expectation that further evidence can lead to progress in identifying cognitive values. Yet, in order to make the contribution of cognitive values empirically assessable, one needs to proceed on the basis of specific research goals. Such goals can be cognitive and non-cognitive at the same time – a status then transferred to the derived values.[4]

3.2 The Problem of Blind Spots

Presuming a distinction between cognitive and other values, AA claims that if scientists stay agnostic whenever epistemic theory assessment cannot be settled without invoking non-cognitive values, the value-freedom of this assessment will be preserved. Even if the cognitive/non-cognitive-distinction is granted, there remains a further problem: Non-cognitive values might decisively shape the research before we even get to the question of how to assess certain theories.

The current VFI explicitly allows for non-cognitive values shaping decisions relevant to the contexts of discovery and theory pursuit. It considers the assessment of theories (in other terms, their justification) to be independent of these contexts. To see why this presumption is wrong, it is first of all important to realize that respective decisions amount to more then the initial selection of a topic; they extend to choices about what research gets funded, which hypotheses are generated, which variables are considered most relevant to their assessment, which data are accordingly generated, how much initial plausibility is assigned to hypotheses, which are considered worthy of extensive testing and continuing effort in the face of anomalies, etc.

[4]For a discussion of whether this dependence of cognitive values on specific research goals leads to a situation of incommensurability of research done in different contexts and with different aims, cf. Bueter (2010).

Secondly, and crucially, these choices do have an epistemic impact. Value-influences on these decisions are transferred into theory assessment via the question, which theories get developed and pursued (and which not) and which data are accordingly generated (and which not). According to VFI, a theory is evaluated in terms of its empirical adequacy and, especially in a situation of competing empirically adequate theories, its exhibition of cognitive values. As pointed out earlier, the underlying hypothetico-deductivism implies that theories and empirical evidence do not stand in a relation of direct logical consequence. This means that it is not always clear which empirical data are essential for a theory's evaluation and what aspects are most significant. Which data a theory needs to account for in order to be considered empirically adequate then depends on selections and ascriptions of significance made by the research community. Especially when it comes to the question what data are already given that a theory needs to explain or be at least compatible with, this makes theory assessment dependent on choices made by the research community at earlier stages of the research process. The same holds for the question of whether and what theoretical rivals have been developed so far that a theory has to outplay in terms of cognitive values.[5]

Only because a theory is empirically adequate and exhibiting cognitive values, it's acceptance is therefore not necessarily value-free. Its acceptability may in part stem from a lack of alternatives. At issue here is not a deliberate ignorance of rival accounts or problematic data, but rather unconscious blind spots. These blind spots can be value-laden, when value-laden background beliefs lead to the invisibility or apparent insignificance of alternative theories or data. This argument is made poignantly by Elliott and McKaughan:

> The degree of evidential support for a theory clearly depends both on the array of available theories and on the set of data at hand. Therefore, to the extent that the nonepistemic values associated with discovery and pursuit influence the available theory and data, they affect theory appraisal. (Elliott and McKaughan 2009, 600)

Such an indirect value-ladenness of theory assessment via value-laden spots in discovery and pursuit cannot be eliminated by adherence to the value-free ideal. Even if theory assessment is based exclusively on cognitive values and empirical adequacy, this does not foreclose the possibility that something has been overlooked – and a theory's status might be judged otherwise in the light of further data and alternatives, if they would exist. Moreover, this problem cannot be solved by AA since blind spots shaping earlier stages of research will not always be eliminated by waiting for further evidence. That would mean that we would have to wait for completeness in terms of data and theoretical alternatives. Yet, such a completeness would, even if it were in principle attainable (which is actually highly doubtful), be undecidable: There just cannot be any guarantee that no alternatives or important aspects were missed. In consequence, the very choice between theories

[5]This point has been made first by Ohkrulik (1994), who argues that the selection of the best a number of sexist theories according to cognitive values does not make the best candidate value-free. Instead, its comparative status just stems from a lack of non-sexist alternatives.

we are presented with can be value-laden – and that kind of value-ladenness cannot reasonably be expected to be eliminated by staying agnostic instead of making a choice.[6]

In principle, there seem to be two different ways to react to the problem of possible blind spots: purity or pluralism. A defence in terms of purity would extend the exclusion of non-cognitive values to the contexts of discovery and to pursuit, arguing that agenda-setting and ascriptions of significance should be guided by cognitive values only, too. This would not eliminate the possibility of blind spots, but it might be said to prevent blind spots that conform to value-laden background beliefs (e.g., invisibilities caused by shared sexist or racist beliefs). However, this option confronts a number of problems. For example, Kitcher (2001) argues that, first, ascriptions of significance are inevitable, if science is to be effective at all and not to be caught up in the gathering of countless trivial truths. Such significance ascriptions can very well be epistemically motivated, if they are made in terms of what research seems scientifically promising, important, or fruitful. Yet, such epistemic significance hinges on the current state of knowledge and the problems it reveals. Since this state is the result of an historical development that has been continuously shaped by social and practical concerns, the exclusion of such factors at this point would not make significance ascriptions value-free (cf. Kitcher 2001. ch. 6 and 7). Instead, a retreat to purity would simply amount to the perpetuation of earlier blind spots and value-influences, not to value-freedom. Moreover, to opt for purity here would imply a rejection of research goals that are cognitive *and* non-cognitive as illegitimate. Presuming the argument from 3.1, such a rejection would make the justification of certain values as cognitive (via goal-conduciveness) much harder, thereby undercutting its own foundation.

Therefore, the better solution here is to opt for a pluralism of values and goals in the research community in order to detect blind spots. It is unavoidable that values play a role in the context of discovery and theory pursuit – so it is important to avoid one-sidedness and enhance the diversity of perspectives on a given research field. Such pluralism will not give any guarantee against blind spots either; yet, it is the best option around.

4 Conclusion

To summarize, one difficulty for AA is that it presumes a context-independent distinction between cognitive and non-cognitive values, which is problematic and unlikely to be advanced by further evidence – wherefore waiting and staying agnostic is of little help. Moreover, even if one grants this distinction, the argument from blind spots is fatal for the agnosticism defence of the VFI. Here as well, waiting

[6]Cf. also Bueter (2015) for a more detailed discussion as well as an illustration of the argument from blind spots by a case study on women's health research.

for further evidence will not do the trick of securing value-freedom, since there cannot be any guarantee that nothing was missed. AA holds that the value-freedom of science can in principle be maintained if problematic decisions on theory choice are suspended. Yet, the very choice we are presented with as well as the criteria guiding our judgment can already be value-laden.

AA would only be a successful defence of VFI in the face of underdetermination if waiting for more evidence could be expected to overcome these problems. However, I have argued that neither the establishment of cognitive values nor the minimization of blind spots in discovery are tasks that can be reduced to gathering more empirical data. In order to proceed at all, we have to make decisions on research goals as well as on significance and insignificance of aspects of the world. These decisions can be value-laden and they cannot all be postponed. Postponing comes with considerable epistemic costs, as Biddle says – costs so high that actually not much is left. Moreover, the problem at hand refers to epistemic evaluation and thus cannot be overcome by distinguishing acceptance for epistemic from acceptance for practical reasons. It is not only the pressure of practice forcing us to make value-laden decisions; instead, the options and criteria we have regarding those practical as well as epistemic decisions are already affected by non-cognitive values. In the last consequence, this argument does not only refute AA, but also VFI. VFI cannot fulfil its function of ensuring the value-freedom of theory assessment, since even complete adherence to its standards (i.e., exclusive reference to empirical adequacy and cognitive values) can leave us with value-laden results.

References

Betz, G. (2013). In defence of the value-free ideal. *European Journal for Philosophy of Science, 3*, 207–220.

Biddle, J. (2013). State of the field: Transient underdetermination and values in science. *Studies in History and Philosophy of Science, 44*, 124–133.

Bueter, A. (2010). Social objectivity and the problem of local epistemologies. *Analyse and Kritik, 32*, 213–230.

Bueter, A. (2015). The irreducibility of value-freedom to theory assessment. *Studies in History and Philosophy of Science, 49*, 18–26.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science, 67*, 559–579.

Elliott, K. C., & McKaughan, D. J. (2009). How values in discovery and pursuit alter theory appraisal. *Philosophy of Science, 76*(Proceedings), 598–611.

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science, 81*, 1–21.

Elliott, K. C., & Willmes, D. (2013). Cognitive attitudes and values in science. *Philosophy of Science, 80*(Proceedings), 807–817.

Giere, R. N. (2003). A new program for philosophy of science? *Philosophy of Science, 70*, 15–21.

Haack, S. (1996). Science as social? – yes and no. In L. Hankinson Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 79–83). Dordrecht: Kluwer.

Harding, S. (1992). After the neutrality ideal: Science, politics, and "strong objectivity". *Social Research, 59*, 567–587.

Howard, D. (2009). Better red than dead – Putting an end to the social irrelevance of postwar philosophy of science. *Science and Education, 18*, 199–220.

Jeffrey, R. C. (1956). Valuation and acceptance of scientific hypotheses. *Philosophy of Science, 23*, 237–246.

Kitcher, P. (2001). *Science, truth, and democracy*. Oxford: Oxford University Press.

Kourany, J. A. (2003a). A philosophy of science for the twenty-first century. *Philosophy of Science, 70*, 1–14.

Kourany, J. A. (2003b). Reply to Giere. *Philosophy of Science, 70*, 22–26.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.), *The essential tension* (pp. 320–339). Chicago: University of Chicago Press.

Laudan, L. (1977). *Progress and its problems*. London: Routledge & Kegan Paul.

Laudan, L. (2004). The epistemic, the cognitive, and the social. In P. Machamer & G. Wolters (Eds.), *Science, values, and objectivity* (pp. 14–23). Pittsburgh: University of Pittsburgh Press.

Levi, I. (1960). Must the scientist make value judgments? *The Journal of Philosophy, 57*, 345–357.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton: Princeton University Press.

Longino, H. E. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 39–58). Dordrecht: Kluwer.

McMullin, E. (1983). Values in science. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1982* (Part II: Symposia, pp. 3–28). Chicago: University of Chicago Press.

Mitchell, S. (2004). The prescribed and proscribed values in science policy. In P. Machamer & G. Wolters (Eds.), *Science, values, and objectivity* (pp. 245–255). Pittsburgh: University of Pittsburgh Press.

Ohkrulik, K. (1994). Gender and the biological sciences. In M. Curd, & J. A. Cover (Eds.), (1998), *Philosophy of science: The central issues* (pp. 192–208). New York: Norton.

Rooney, P. (1992). On values in science: Is the epistemic/non-epistemic distinction useful? In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1992* (Part 1: Contributed Papers, pp. 13–22). Chicago: University of Chicago Press.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science, 20*, 1–6.

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science, 40*, 92–101.

Part IV
Causality

Learning About Constitutive Relations

Lena Kästner

1 Interlevel Experiments

Contemporary cognitive neuroscience aims to discover how biological brains give rise to the mental, psychological or *cognitive* phenomena. Scientists' attempts to explain these phenomena by recourse to empirical studies of brain activity have recently gained attention in philosophy of science (e.g. Bechtel 2008; Bickle 2006; Craver 2007b; Gillett 2011; Leuridan 2012; Machamer et al. 2000). The *mechanistic view* is particularly prominent. Its proponents hold that to explain a phenomenon means to identify its underlying *mechanism*, viz. "a set of entities and activities organized such that they exhibit the phenomenon to be explained." (Craver 2007b, p. 5) That is to say, we must find the components or *constitutively relevant parts* of the phenomenon's implementing mechanism and how they *causally* work together. But how can this be done?

Carl Craver (2007a,b) advocates a *manipulability-based* assessment of causal relations among a mechanism's parts as well as of constitutive relations between a phenomenon as a whole and each of its components. A mechanistic component, according to Craver, is a spatial part of the mechanism in question that is constitutively relevant to the operation of the mechanism as a whole in the sense that manipulating the component will affect the behavior of the mechanism as a whole *and* vice versa (cf. Craver 2007b, pp. 152–153). Such constitutive relations between a mechanism and individual components can be tested in *interlevel* studies, viz. studies where a manipulation and the detection of its effect happen at

L. Kästner (✉)
Berlin School of Mind and Brain, Humboldt Universität zu Berlin, Berlin, Germany
e-mail: mail@lenakaestner.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_12

different *levels* of mechanisms.[1] Causal relations among mechanistic components, by contrast, are tested in *intralevel* studies, viz. studies where manipulation and detection techniques are applied at the same mechanistic level (cf. Craver 2007a, p. 12). According to Craver's *mutual manipulability criterion* (MuMa), we have *sufficient* evidence to infer a *constitutive* relation if we can manipulate, or *wiggle*, the (upper-level) mechanism as a whole through manipulation of a (lower-level) component in a bottom-up experiment *and* manipulate, or *wiggle*, that (lower level) component through manipulation of the (upper-level) mechanism as a whole in a top-down experiment.[2] Craver explicitly borrows this idea of *wiggling* something to see what else will *wiggle along* from Woodward (2002, 2003). According to Woodward's celebrated *interventionist theory of causation* interventions uncover causal (explanatory) relevance relations. Put briefly, Woodward's idea is that I is an intervention on some factor X with respect to some other factor Y if and only if (i) I is causally relevant to X, (ii) I overrides all other (potential) causes of X, (iii) I is causally relevant to Y through X only, and (iv) I is independent of any other causes of Y (cf. Woodward 2003, p. 98). Since Woodward's view nicely captures certain well-known principles of experimental design such as ideas about screening off and controlling for interfering factors, his interventions feature prominently in contemporary accounts of scientific explanation, especially in the context of mechanisms.

That scientists apply manipulations to investigate and eventually explain phenomena is uncontroversial. In cognitive neuroscience, interlevel manipulations are particularly prominent as cognitive phenomena are commonly believed to be implemented by neural processes in the brain.[3] A prime example in this context is hippocampal long-term potentiation (LTP) as the cellular mechanism for spatial memory (e.g. Bickle 2006; Craver 2007b; Kandel 2006). In accordance with MuMa, we find a mutual manipulability relation here: interfering with LTP disrupts spatial memory capacities and spatial memory formation is accompanied by hippocampal LTP. Similar success stories of interlevel experimentation can be found all over cognitive neuroscience. They even include much more complex higher-level phenomena such as, e.g., language processing (see e.g. Friederici

[1] On Craver's view, mechanisms are organized in nested hierarchies. "Lower levels in this hierarchy are the components in mechanisms for the phenomena at higher levels." (Craver 2007b, p. 170) Since the hierarchical relationship between levels is based on componential relations, it can only be determined within any one mechanism. We have no way of saying that two independent mechanisms are at the same or different levels (see also Fazekas and Kertész 2011).

[2] The crucial difference between bottom-up and top-down experiments is where the manipulation is being applied and where its effects are being detected. In a top-down experiment, we manipulate at the top and assess the changes this induces at the bottom level. In bottom-up experiment, we induce a manipulation at the bottom and observe the changes it elicits at the top.

[3] The background assumption here is, of course, that cognitive phenomena are *at a higher level* than neural or brain processes that implement, realize, or otherwise *give rise to* them—an intuition hardly ever debated. Note that these general levels are different from mechanistic levels (see footnote 1). For current purposes we can gloss over this, however.

2002; Hickok and Poeppel 2007; Patterson et al. 2007). Admittedly, scientists' explanations of language processing do not go 'all the way down' to cellular and molecular processes. Yet, they are mechanistic in character: they identify entities and activities organized such that the phenomenon in question is exhibited.

Though appealing given what we see in scientific practice, the 'mechanisms and interventions'-view remains unsatisfying for several reasons. Strictly speaking, we cannot actually carry out Woodwardian interventions in interlevel contexts—at least not as long as we stick to Woodward's original definition of interventions. Put in a nutshell this is because Woodward's interventionism is not suited for handling variables that stand in non-causal dependence relations. However, if mechanisms are constituted by their components, they non-causally depend on them.[4] But even assuming that we *can* employ Woodwardian interventions in interlevel contexts, interventions alone will hardly get off the ground: we can only intervene into components or observe an intervention's effects on components once we know what the (potential) components are. And this is not usually something we can find by intervention studies alone. In fact, it seems that interventions are something we will usually see somewhere in the middle of the discovery process rather than at the very beginning (if we see them at all). Similarly, there are many cases in empirical research where the desired interventions cannot be carried out. And even if they can be performed, observed manipulability will underdetermine the precise dependence relation at hand. Luckily, science is not limited to interventions. To highlight the role that non-interventionist manipulations play is my objective in this paper.

I will begin by discussing the limits of interventionist manipulations in more detail in Sect. 2. This is not to deny that interventions are an important part of scientific practice, of course. It is just to highlight that if we want a full understanding of how scientific explanations are constructed, we need to step beyond the by now familiar 'mechanisms and interventions'-picture. In Sect. 3, I will argue that this can be achieved—at least to a certain extent—by experimental manipulations that do *not* qualify as interventions in Woodward's sense. I shall call these *mere interactions*. The distinction between mere interactions and interventions will be explicated in Sect. 4. My conclusion will be that understanding how scientists explain phenomena, and specifically how they learn about constitutive relations, involves thinking beyond interventions. While I do agree that Woodward's interventionism is a plausible starting point for understanding how explanatory information can be gathered by experimentation, it is clearly too short-sighted as a general account of how scientific explanations are constructed.

[4]Put in a nutshell, the problem is that in order for Woodward's definition of an intervention to be satisfied, the variables under consideration must *not* be non-causally related. The question of whether or not interventions can be usefully applied where non-causal dependence relations hold makes for a whole other debate both in the context of mechanistic explanations and mental causation (see e.g. Baumgartner 2010; Gebharter and Baumgartner 2013; Hoffmann-Kolss 2014; Williamson 2013; Woodward 2008, 2011, 2013). Since the point I am making in the current paper is quite independent, this is debate not at issue here (but see Kästner 2014).

2 The Limits of Wiggling

If we want to use manipulability to assess causal and/or constitutive dependence relations as mechanists suggest, we must first know what we are targeting with our manipulations, how to target it, and where and how to assess the effects. In the context of mechanistic explanations this means that in order to apply MuMa we must know what phenomenon we are trying to explain, what a possible implementing mechanism could be, what its potential components are, and how to manipulate both the mechanism as a whole and individual potential components. Put in interventionist terms, that is to say we must know what the different variables (representing a mechanism and its potential components) are and how to measure and manipulate their values. It will further be advisable to know whether our variables are at different mechanistic levels. For intralevel relations among mechanistic components are supposedly causal in character while mechanistic interlevel relations are explicitly characterized as *non-causal* dependence relations (e.g. Craver 2007a; Craver and Bechtel 2007). According to MuMa we have sufficient evidence to assume a constitutive relation where we find *interlevel* mutual manipulability. But to identify interlevel manipulability relations we first have to know whether a given manipulation is an interlevel manipulation to begin with. And even if we know that, MuMa remains limited. It is only a *sufficient* not a necessary condition for identifying mechanistic constitutive relations, after all. As such it is informative only for cases where there is either manipulability or non-manipulability in *both* directions, viz. bottom-up and top-down. Thus, MuMa does not say anything about intermediate cases where there is manipulability in one direction only (see Craver 2007b, p. 17).

But cases where there is manipulability in one direction only are abundant in empirical research. This might be, e.g., due to methodological limitations or for ethical reasons. For illustration consider research on language processing. We may carry out lots of top-down studies where we have participants engage in language processing tasks and observe the effects on cortical activations. Despite limited temporal and/or spatial resolution, such top-down studies have significantly advanced our understanding of how the brain processes natural language (see e.g. Friederici 2002; Hickok and Poeppel 2007). Complementary bottom-up studies would require us to interfere cortical language processing in regions of interest like, say, superior temporal gyrus (STG) which is known to be involved in phonological processing. However, lesion studies in humans are considered unethical and animal models are not available for studying language processing. In our case neuroscientists may resort to transcranial magnetic stimulation (TMS). TMS temporarily disrupts neural processing in stimulated areas allowing experimenters to create a *virtual* lesion. While we may apply TMS to STG, its reach is limited to superficial cortical structures (plus localization is somewhat tricky anyway). Therefore, we cannot use TMS to target deeper cortical structures (e.g. hippocampus) that may be regions of interest (e.g. in the context of memory research). Where complementary bottom-up and top-down studies are unavailable, problematically, we cannot find mutual

manipulability and may be lured into giving the unidirectional manipulability we observe a causal interpretation. But if we are investigating mechanistic interlevel relations—which are supposed to be constitutive and not causal—we need to avoid this. In order to do so, we probably will have to rely on additional evidence, e.g. evidence about what is at the same and different levels. Once we know this, we may tease causal and constitutive relations apart.

The same principle problem occurs where complementary bottom-up and top-down studies have to resort to different experimental tools and settings. We may study phonological processing bottom-up using TMS, but to study it top-down we most likely need to employ a neuroimaging technique such as electroencephalography (EEG) or functional magnetic resonance imaging (fMRI). For instance, we could give subjects a phonological processing task and see how this elicits or increases activation in certain cortical areas—such as STG—compared to rest.[5] Taken individually, each of these experiments invites a causal interpretation: phonological processing *elicits* STG activations, 'lesioning' STG *disrupts* phonological processing. The only way to foreclose this confusion is to know that we are manipulating across levels rather than at the same level. And to know that we must not only know what the variables are, how to manipulate and measure them, but also which variables are at the same level. In the mechanistic context, that is to say we must know what a phenomenon's (potential) implementing mechanism is and what its parts (and thus potential components) are. With this knowledge about part-whole relations, we can also address a second problem for MuMa: we can settle in which direction a constitutive relation obtains. When we find mutual manipulability, MuMa says, we can infer a constitutive relation. But if A is manipulable by intervention into B and B is manipulable by intervention into A this leaves underdetermined whether A constitutes B or B constitutes A. Now if we know that B is a part of A, we can conclude that B constitutes A; for parts constitute their wholes and not vice versa.

Taken together, this makes it quite plain that we need to supplement MuMa—and interventions more generally—with knowledge about what a phenomenon's (potential) implementing mechanism as well as what its (potential) components are. Of course, none of this is to say that Woodwardian interventions and MuMa are of no use in science. It is merely to point out that they are limited and can only be part of the story when it comes to understanding how scientists explain phenomena.[6] Another, usually earlier, part of this story has to be about how scientists identify phenomena, their implementing mechanisms and the potential components of these mechanisms. This takes us to the next section.

[5]Notice again, that the picture I sketch here is highly simplified and idealized. But for current purposes it shall suffice.

[6]I am not saying either that it is *only* Woodwardian interventions that are limited. In empirical research, our tools—interventionist or not—are almost always imperfect.

3 Beyond Mutual Manipulability

How can we find mechanisms and their potential components without having to rely on assumptions about levels and componential relations to begin with? The solution is as simple as this: remove your interventionist glasses and look at scientific practice more closely. Once freed from the focus on interventions, we will soon realize that thinking in terms of interventions only is much too short-sighted. Though interventions are an important part of scientific practice, they are certainly not all there is. And they are not how the business usually gets started. For we need to know about mechanisms and their parts before we can target them with interventions.

But how do we get started if we want to explain a phenomenon, identify its underlying mechanism, its parts, and how they work together? The perhaps simplest way to do this is *passive observation*. We can observe a phenomenon to delineate it from things happening in its surroundings. When we have isolated the phenomenon of interest we can start looking for its implementing mechanism. For illustration consider the simple case of a fridge. First, we find that cooling occurs at the inside of the fridge when it is closed and plugged in. We have thus delineated the phenomenon and found the machine that produces it. Now if we want to explain the fridge's cooling mechanistically, we must look for the parts of the fridge, how they are organized, and how they work together. How do we do this? We could just look behind the fridge's backplane and observe what the components are, what they do, how they are arranged and interlinked.[7] But often we do not see much; passive observation is rather limited. When it comes to neural processes, for instance, we cannot just watch what is happening by unaided eye.

That "the secrets of nature reveal themselves more readily under the vexations of art than when they go their own way" (1620, XCX) was already recognized by Francis Bacon. Merely observing an object or phenomenon, that is, is not as telling as inspecting it under manipulation. This principle is the very backbone of empirical research: when trying to understand and explain how a phenomenon comes about, scientists systematically manipulate it. Such manipulations *can* be interventions, of course: we can *wiggle* something to see if/how something else *wiggles* along. But there is more to experimental manipulation than that. If we cannot just watch to learn how the fridge works, we can take it apart to identify its parts and try and keep track of how they are linked to understand how the different parts work together to produce the fridge's cooling. This straightforward spatial decomposition clearly is a manipulation. After all, we do change the fridge substantially as we remove part after part from its proper arrangement, we disrupt its cooling capacity, and so on. But this kind of manipulation is clearly not an intervention in Woodward's sense: we do not decompose the fridge to see what happens if we decompose a fridge. We

[7] Subsequently we can, of course, apply interventions (even MuMa) to test for the *constitutive relevance* of various parts.

know we will reveal its spatial parts, and that is why we do it. Thus, decomposition of this kind is a non-interventionist manipulation.[8]

There are many different varieties of such non-interventionist manipulations. I shall collectively refer to them as *mere interactions*. The major difference between interventions and mere interactions is how what is being manipulated is related to what is being studied. With interventions, we manipulate X to see how it affects Y. With mere interactions, we manipulate X to learn about features of X itself, e.g. what separates it from its surroundings, how it is structured, and what its parts are. Like interventions, mere interactions typically change the system or phenomenon of interest in some way. But unlike interventions, mere interactions are *not* applied in order to see what their effects are going to be. Scientists know what the effects of mere interactions will be. That is why they purposefully employ them as tools. To illustrate the distinction consider another example. Suppose you are given a sample of neural tissue. For simplicity's sake assume that the sample only contains a network of interconnected excitatory neurons where all synapses work flawlessly and nothing interferes with proper signal transduction.[9] We know certain basic neurophysiological facts: that action potentials are generated in the cell body once the neuron's membrane depolarizes above a certain threshold, that these action potentials are propagated along axons and received by another neuron's dendrites, and that the receiving neuron depolarizes in response to the incoming signal. Given that, we know that action potentials will usually be propagated between neurons that are anatomically connected in the right way (i.e. through axon and dendrite). Now suppose we want to find out about the anatomical connections among neurons in our sample. Which neurons are linked to one another and between which neurons in our sample will electrical signals be propagated?

One strategy to approach this question is to *wiggle* a neuron at one point and see if a neuron at the other *wiggles along*. In order to do this, we can induce a current (I) into the first neuron, measure its membrane potential (M_1) and also measure the second neuron's membrane potential (M_2) where both M_1 and M_2 can take values 0 or 1 if they are below or above threshold, respectively. This is a classical intervention study: I qualifies as a Woodwardian intervention on M_1 with respect to M_2 (see Fig. 1). If I sets the value of M_1 from 0 to 1, the neuron will—given appropriate

$$I \longrightarrow M_1 \longrightarrow M_2$$

Fig. 1 Woodward-style intervention I on M_1 with respect to M_2

[8]This is not to say, of course, that we cannot study e.g. the fridge's electrical circuit using interventions. But I want to raise attention to the fact that *before* we can employ interventionist manipulations we usually need to figure out what we can manipulate, i.e. find the (potential) components in a mechanism. How these work together may best be studied by interventions, but that is another question.

[9]This is obviously highly idealized and oversimplified. But for illustrative purposes this toy example will do.

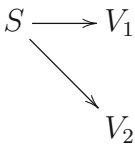


Fig. 2 *S* is *not* a Woodward-style intervention on V_1 with respect to V_2

background conditions—fire. Now if we observe that once the first neuron fires the second neuron fires, too (i.e. M_2 changes from 0 to 1), we can infer—given our basic knowledge of neurophysiology—that there exists an anatomical connection between the two neurons along which the induced signal is propagated. If we repeat this procedure over and over again, we may eventually reconstruct the entire neural network in the tissue sample.

But there is another way to achieve this: we can stain the tissue sample (*S*), e.g. by using *Golgi's method*. Bathing the sample in a potassium dichromate solution will induce a chemical process known as *black reaction* in the neurons which makes otherwise opaque cell membranes visible. We may conceptualize that as changing the values of variables V_1 and V_2 standing for the visibility of the membranes of neurons from values 0 (not visible) to 1 (visible) (see Fig. 2). For simplicity, assume that our staining technique works reliably and flawlessly and does not miss out on anything such that *S* is guaranteed to reveal all anatomical connections between neurons in the sample. Once the overall network is visualized we can immediately see what the anatomical connections are—and thus infer between which neurons an electrical signal will be propagated.[10]

Now one might argue that *S* still is an intervention on V_1 just as *I* is an intervention on M_1. However, the crucial difference between *S* and *I* is that while *I* is an intervention on M_1 *with respect to* M_2, *S* is *not* an intervention on V_1 with respect to V_2 (or vice versa). Since *S* acts as a common cause of both V_1 and V_2, it cannot, according to Woodward's definition, be an intervention on either V_1 or V_2 with respect to the other.[11]

In the first case, we manipulated M_1 to see how this affects M_2. It was thus illuminating, albeit maybe hypothesized, to see a change in M_2 resulting from *I*. This

[10]As an anonymous reviewer has rightly noted, the first experiment uses causal information (about electrical signal transmission) to infer anatomical structures, while the second experiment employs anatomical information to draw inferences about causal links. However, this does not mean the experiments had actually different aims. They merely used different research strategies and different aspects of available background knowledge to investigate the same thing, viz. layout of the neural network in the tissue sample. This is a feature highly characteristic of empirical research practice.

[11]This violates condition (iii) of Woodward's definition of an intervention (cf. Woodward 2003, p. 98). The problem is not per se that *S* is a fat-handed intervention (i.e. an intervention that causes many variables to change at once) but that it is precisely the relation between the co-varying variables that is at issue. It is possible, of course, that *S* is an intervention on V_1 or V_2 with respect to some other variable representing an effects of V_1 or V_2, respectively. However, that would be quite a different scenario.

is what makes *I* an intervention on M_1 with respect to M_2. We looked for whether M_2 *wiggled* along when we *wiggled* M_1. *S*, on the other hand, is not an intervention on V_1 with respect to V_2; it immediately and independently affected both V_1 and V_2. We knew *S* would affect both V_1 and V_2 in the same way, this is precisely *why* we stained the cells: we wanted to be able to see them. What is illuminating about *S* is not the fact *that* the staining made the membranes visible but the overall anatomical structure it reveals. Although they have been there all along, it is only after the black reaction that we learn about the anatomical connections in the sample. This is why I call manipulations like *S mere interactions*: we do not elicit or interfere with the very thing we are studying but *merely* make it accessible to our observation—albeit maybe employing sophisticated tools and methodologies to help the unaided eye. When we use mere interactions, our aim is *not* to understand the connection between a manipulation and its effect (we already know that). Rather than trying to elicit or interfere with a phenomenon our aim is to observe something that is relatively independent of the manipulation as such but that only becomes accessible once the manipulation is applied.

4 Interventions and Mere Interactions

Clearly, both "intervention" and "mere interaction" are causal notions. Both refer to types of manipulations that have some kind of effect on the investigated phenomenon or system. Both *I* and *S* do, after all, influence on the tissue sample. My distinction between interventions and mere interactions takes issue with the precise relation between what it is that is being manipulated and what it is that is being studied. While *I* elicits the neural signal propagation from the first to the second neuron by manipulating the first, *S* targets the sample as a whole to *merely* reveal structural information without changing anything about it. While Woodward-style *interventions* usually are manipulations *with respect to* the explanandum phenomenon, mere interactions do not typically affect it. They often reveal spatial parts (i.e. mechanisms' potential components) and their organization rather than wiggling them.

Mere interactions can be any kind of procedure or manipulation that makes accessible aspects of the system or phenomenon in question that are inaccessible without it. This may include decomposition, microscopy, staining, and various imaging techniques as well as employing different kinds of measuring devices. Note that mere interactions are importantly different from passive observation. They are, like interventions, still a species of *manipulation*—they involve a researcher applying tools for a means rather than just passively observing. While there may be quite a wide variety of mere interactions, all of them are united by their revealing e.g. organizational features of a system or information about components. As such, mere interactions do make a vital contribution to the explanatory enterprise—one that is, as the toy examples above have demonstrated, neither captured by interventions nor by passive observation.

Mere interactions will often provide us with the very knowledge required to carry out interventions to begin with, especially where this knowledge is unavailable through passive observation alone (for instance, just looking at the tissue sample is not enough if we want to learn about anatomical structures). Thus, mere interactions sometimes are a *prerequisite* for interventions. Consider the tissue sample again. How do we know where to induce the current and where to measure the membrane potentials? In practice, current induction would usually be preceded by some kind of visualization such that individual cells can be targeted and effects on individual cells can be measured. Perhaps we could also find cells in the sample by trial and error using interventions. Perhaps we could even reconstruct the whole network in the tissue sample this way. But it would be a tedious thing to do. It is much more efficient to stain the tissue sample and reveal the whole network of neurons in one go. This is typical for mere interactions: that they reveal a lot of information at once which subsequent intervention studies can draw on. We can use mere interactions to learn about part-whole (and thus potentially constitutive) relations rather than having to know about them before we start manipulating (as is the case with interventions). As such, mere interactions can be applied in cases where interventions are inapt or even impossible.

Note that my distinction between interventions and mere interactions is independent of whether a manipulation involves actually touching something (for a distinction along these lines see Silva et al. 2013). An intervention may consist in removing a chunk of brain by a surgeon, but it could also be a toxic lesion in which case nobody ever touches the manipulated brain. Analogously, a mere interaction may involve slicing tissue and staining cells, but it could also be positioning an instrument for measurement without touching what is being measured. Likewise, I am not concerned with the distinction between manipulations of actual physical systems and simulation experiments (as Morgan (2003) suggests) or modeling (as Scholl and Räz (2013) suggest). Whether we manipulate a model, a simulation, or a "real" thing is irrelevant for my distinction between mere interactions and Woodward-style interventions. Also, my distinction is independent of whether we manipulate an entity, its activity, or the organization of the whole system and whether the manipulations we consider are interlevel or intralvel manipulations. Neither does my distinction depend on the actual causal and/or constitutive structure of the investigated system or the actual (causal) process of manipulation.

In fact, whether a manipulation qualifies as intervention or mere interaction is somewhat relative to our driving research question: the same manipulation on the same structure can be interpreted as an intervention or a mere interaction *depending on what use we put it to* and which of its effects we are interested in. Recall the staining case again. We wondered whether an electrical signal would be propagated between two points in our sample. Given our background knowledge (and idealizing assumptions) we knew that our answer could be based on anatomy: if there is anatomical connection between the neurons in question the signal will be propagated. We thus applied *S* as a tool to visualize the anatomy in the tissue sample. *S* did not change anything about the anatomy—it was a mere interaction, *not* an intervention—, yet *S* enabled us to answer our driving research question.

Now suppose we use the same tissue sample and the same staining technique but are wondering what color the cell membranes will assume. We apply the stain in order to see what the cells will look like. In this setting, *S* does affect the phenomenon of interest, viz. the cell membranes' assuming a certain color, and thus qualifies as an intervention on the neurons with respect to the color of their membranes. This context dependency is not my invention. It is already evident in Woodward's characterization of interventionist explanations as *contrastive*: whether a manipulation is an intervention essentially depends on whether we are *interested in the difference* (or contrast) *it produces*. If we wiggle something to see what else will wiggle along we are clearly interested in the effects our manipulation brings about—we intervene. But if we use manipulations as tools, and already know what their effects will be, we merely interact. And it is this merely interacting that gets us what we are looking for: knowledge about mechanisms, their components, and how these are organized. We learn about constitutive relations by mere interactions.

5 Conclusions

Both mechanistic explanations and interventionism capture some interesting and important aspects of scientific practice. Yet, focusing on MuMa and interventions alone is much too short-sighted if we want to understand how scientists explain phenomena. For MuMa—and interventions more generally—to get off the ground, we first need to know what we are manipulating, i.e. what the variables are or, in a mechanistic context, what a mechanism's potential components are. Since components are constitutively relevant parts, a good strategy will be to first look for spatial parts and subsequently apply MuMa. Knowing about parthood relations will further enable us to establish the direction in which a possible constitutive relation obtains. But how do find out about a mechanism's spatial parts?

This is not typically done by interventions. Passive observation is one way to gather parthood information but it is often quite limited. Besides, scientists often use manipulative non-interventionist strategies in this context. To capture this, I introduced to concept of *mere interactions*. Mere interactions are manipulations which are best understood as tools enabling observations otherwise impossible. Embracing mere interactions, we can supplement interventions and MuMa without being restricted to passive observation only. Without them, the familiar 'mechanisms and interventions'-picture misses out on something crucial that is happening early on in the discovery process.

Though both "intervention" and "mere interaction" refer to some kind of manipulation, mere interactions do not qualify as interventions in Woodward's sense. Unlike interventions, which we employ to learn about their effects, we utilize mere interactions as tools *because we know what their effects will be*. Though their effects are known, mere interactions are important, sometimes even vital, to scientific practice as they can make available (i) information otherwise hidden from our view, (ii) information required to carry out subsequent intervention studies, and

(iii) information required to apply MuMa and to disambiguate which dependence relation underlies observed manipulability.

Mere interactions are actually all over the place in science. For examples in cognitive neuroscience just consider Felleman and Van Essen's (1991) work on the visual system (based on histology), or research into neural oscillation patterns (e.g. Buzsaki and Draguhn 2004; Düzel et al. 2010). Though cutting up a brain according to its anatomy and planting electrodes on participants' scalps clearly are manipulations, they do not usually qualify as interventions. These manipulations are—just like staining a tissue sample to reveal the overall architecture of a neural network—mere interactions. Still, such manipulations play an important role in the construction of many scientific explanations.[12]

Philosophers' recent focus on interventions has lead them to overlook the crucial role that non-interventionist manipulations play in empirical science. Yet, thinking beyond interventions bears significant potential for philosophy of science. For one thing, mere interactions can supplement intervention-based accounts of scientific explanations. They give us a means to identify (potential) mechanisms and their potentially constitutively relevant parts and thus gather the very knowledge that intervention-based research (including the application of MuMa) has to rely on. For another, acknowledging the role that mere interactions play in experimental practice will render accounts of scientific explanation more empirically adequate than a short-sighted focus on interventions only.

References

Bacon, F. (1960 [1620]). *The new organon*. New York: Bobbs-Merrill.

Baumgartner, M. (2010). Interventionism and epiphenomenalism. *Canadian Journal of Philosophy, 40*, 359–384.

Baumgartner, M., & Gebharter, A. (2015). Constitutive relevance, mutual manipulability, and fat-handedness. *British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axv003.

Bechtel, W. (2008). *Mental mechanisms*. London/New York: Routledge.

Bickle, J. (2006). Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese, 151*, 411–434.

Buzsaki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science, 304*, 1926–1929.

Craver, C. F. (2007a). Constitutive explanatory relevance. *Journal of Philosophical Research, 32*, 3–20.

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

Craver, C. F. and Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22*, 547–563.

[12]There is certainly more to be said on different non-interventionist manipulations. It is not only that there are many different varieties of mere interactions. We can further use mere interactions to *pretend* Woodwardian interventions. Unfortunately, this is beyond the scope of this paper; but see Kästner (2014).

Düzel, E., Penny, W., & Burgess, N. (2010). Brain oscillations and memory. *Current Opinion in Neurobiology, 20*, 143–149.
Fazekas, P., & Kertész, G. (2011). Causation at different levels: Tracking the commitments fo mechanistic explanations. *Biology and Philosophy, 26*, 365–383.
Felleman, D. J., & Essen, D. C. V. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*, 1–47.
Friederici, A. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6*, 78–84.
Gillett, C. (2011). Multiply realizing scientific properties and their instances. *Philosophical Psychology, 24*, 727–738.
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*, 393–402.
Hoffmann-Kolss, V. (2014). Interventionism and higher-level causation. *International Studies in the Philosophy of Science*, 28, 49–64.
Kandel, E. R. (2006). *In search of memory: The emergence of a new science of mind*. New York: W. W. Norton & Company.
Kästner, L. (2014). *Philosophy of cognitive neuroscience: Causal explanations, mechanisms & empirical manipulations*. PhD thesis, Ruhr-Universität Bochum.
Leuridan, B. (2012). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *The British Journal for the Philosophy of Science, 63*, 399–427.
Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.
Morgan, M. (2003). Experiments without material intervention. In H. Radder (Ed.), *The philosophy of scientific experimentation* (chap. 11, pp. 216–235). University of Pittsburgh Press, Pittsburgh.
Patterson, K., Nestor, P., & Rogers, T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*, 976–987.
Scholl, R., & Räz, T. (2013). Modeling causal structures: Volterra's struggle and darwin's success. *European Journal of Philosophy of Science, 3*, 115–132.
Silva, A., Bickle, J., & Landreth, A. (2013). *Engineering the next revolution in neuroscience*. New York: Oxford Univesity Press.
Williamson, J. (2013). How can causal explanations explain? *Erkenntnis, 78*, 257–275.
Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science, 69*, S366–S377.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.
Woodward, J. (2008). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: New essays on reduction, explanation, and causation* (pp. 218–262). Oxford: Oxford University Press.
Woodward, J. (2011). Interventionism and causal exclusion. Manuscript. Available at: http://philsci-archive.pitt.edu/8651/
Woodward, J. (2013). Mechanistic explanation: Its scope and limits. *Proceedings of the Aristotelian Society, LXXXVII*, 39–65.

Reconstituting Phenomena

Maria Kronfeldner

1 Causal Complexity and the Trick of the Archer

When many causal factors are involved in the production of a phenotypic trait of an organism, then life scientists call these traits *causally complex*. If the focus is on genetic factors, then such traits are said to be polygenic (rather than monogenic); if environmental factors are also part of the story, then the traits are called multifactorial or even multilevel.

Yet, as a matter of fact, many explanations of such causally complex traits use simplifying strategies in the sense that they ignore most of the respective causal factors. This has led to many of the pernicious nature-nurture wars of the twentieth century with some scientists ignoring genetic factors (nature), and others environmental ones (nurture).

The kind of causal complexity at issue is compatible with the trait itself being of a rather simple nature. Even a trait such as body height – which is simple in the sense that it has no parts and is easy to define and measure (in contrast to a trait such as intelligence or schizophrenia) – is (as far as we know) causally complex. There is not one gene for body height, but lots of genes influencing the trait, each with a tiny effect, and each interacting with lots of other causal factors, including

M. Kronfeldner (✉)
Department of Philosophy, Central European University, Nádor u. 9, 1051 Budapest, Hungary
e-mail: KronfeldnerM@ceu.edu

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_13

environmental ones (see Visscher 2008). The kind of complexity that is at issue in this paper is thus a many-to-one causal complexity.[1]

In principle, we can react to causal complexity of that sort with two simplifying strategies (used intentionally or nonintentionally): (a) by selectively focusing on particular causes, while relegating other causally relevant factors to the status of mere conditions (causal selection); and (b) by dividing the phenomenon that we want to explain into parts or states that are more tractable (reconstituting phenomena).[2] These two simplifying strategies conquer complexity by dividing either the explanans or the explanandum, or both. As a result, we get a partial, simplified picture of, ideally, one-to-one relationships: effects that "have a cause of their own" and causes that "have an effect of their own."

This paper focuses on the second simplifying strategy, reconstituting phenomena, even though the first enters the picture. The paper aims to show that, first, phenomena are moving targets and that, second, there is a mode of reconstitution of phenomena that moves up to a level of greater abstraction rather than down to a level of greater decomposition. Only the latter is currently acknowledged in the philosophical literature discussing complexity and causal explanation in the life sciences, e.g., in Bechtel and Richardson (1993/2000).[3] Part of the reason for this is a focus on mechanistic explanation, which is widespread but not always what scientists intend. Sometimes, they are interested in less – simply in what I will call "difference explanations." A third major aim of the paper is then to derive some consequences regarding the kind of pluralism entailed when scientists reconstitute (i.e., reconstruct) phenomena by moving up to a level of greater abstraction.

Before I proceed to examples of the two modes of reconstituting phenomena, I want to illustrate my first aim of the paper by referring to a Jewish parable, the parable of the king and the archer: There was once a king who was in urgent need of a very good archer. After desperately searching for a time, he discovered a field strewn with bull's-eyes. Inquiring about the archer, he was directed to the local library, where he found Harold reading a book through thick glasses. The king was skeptical and asked Harold how he managed to be such a good archer. Harold replied: "Sir, it is really very simple. I stand in front of the fence and shoot my arrow. Then I paint the target around it, with the arrow at the exact center" (quoted by Schwartz 1998).

Scientists often use the archer's trick: they choose one convenient, easy-to-handle causal factor (e.g., bits of DNA) as an arrow (*causal selection*), shoot it

[1] It can be contrasted with organizational complexity, dynamic complexity, and semantic complexity, the latter, pointing at definitional heterogeneity of the phenomenon and the problems surrounding construct validity.

[2] By "phenomena," I simply mean objects of research, i.e., features of the world that we want to describe and explain. Thus observability is not assumed. On the latter, see the classic paper on the distinction between data and phenomena from Bogen and Woodward (1988). For details on identifying phenomena in actual scientific practice, see Feest (2011).

[3] From whom I take the term "reconstituting phenomena."

(i.e., experiment with it or collect data about it), and then, if necessary, adapt their phenomena (explananda) afterward (*reconstituting phenomena*). Phenomena – the scientist's bull's-eyes – are moving targets that are not fixed once and for all, but are adapted. What is fixed (if at all) is the causal factor as the focus of scientific interest, which itself varies according to disciplinary affiliation. This adaptation of the phenomena, furthermore, does *not* always involve moving *down to a level of greater decomposition* and therefore toward greater proximity of causal relations, i.e., more details. "More detail" is not always what scientists want. Indeed, they might well prefer to *move up to a level of greater abstraction* in order to get rid of those causal factors they want to ignore (given their focus of interest).

In Sects. 2 and 3, I will illustrate the two modes of reconstituting phenomena. In Sect. 4, I will explicate in detail the kind of pragmatic pluralism of causal perspectives that results from the simplifying, complexity-reducing strategy of reconstituting phenomena by moving up to a level of greater abstraction.

2 Reconstituting Phenomena by Moving Down to a Level of Greater Decomposition

To reduce the complexity of causal relations involved in explaining phenotypic traits, scientists often move their explanandum phenomenon down to a level of greater decomposition, i.e., toward greater proximity and increasingly specific one-to-one causal relations. Such a move down to a level of greater decomposition happened, for instance, when scientists adapted their explanandum from the Mendelian picture of "one gene explains one *phenotype*" to the more modest "one gene explains one *enzyme*" paradigm. The enzyme is an organizational part of the phenotype and consequently located at a level of greater decomposition than the respective phenotype; the enzyme is also a more proximate effect of the gene that has the phenotype as a distal effect.

A more contemporary case of this mode of reconstituting phenomena by moving down to a level of greater decomposition, i.e., toward more proximate effects, is the concept of endophenotypes. An explanandum (e.g., the phenotypic trait schizophrenia) gets partitioned into parts, the endophenotypes (e.g., sensory motor gating, oculomotor function, etc.) (Gottesman and Gould 2003). These are more proximate effects of the genes, and the phenotype results from – and is also, in that sense, composed of – them, even though these "parts" are not organizational (mereological) parts of schizophrenia. Furthermore, even though the move to endophenotypes is a move towards parts of the phenotype, it does not involve a *new kind* of explanandum, a new level of abstraction, since endophenotypes are the *same kind of* things as phenotypes and constitute the latter. Yet, the phenomenon (the phenotype as explanandum) is adapted – reconstituted – at a "deeper" level of decomposition.

If scientists remained at the level of the phenotype, despite tremendous causal complexity, then – in the final instance – everything would be connected with everything. But "[i]n pointing at everything, geneticists would point at nothing" as David Goldstein (2009: 1696) recently said.[4] That is why they adapt their explanandum by moving down to a level of greater decomposition, which thereby gives them increasingly proximate and usually also increasingly specific one-to-one causal relations.

Yet scientists can also move *up to a level of greater abstraction*, which is a move to a *new kind of* explanandum phenomenon. In the example I shall use, which is again connected to the phenogenesis of traits, it is a move from explaining *traits* to explaining *trait differences*.

3 Reconstituting Phenomena by Moving Up to a Level of Greater Abstraction

To avoid semantic issues about construct validity (how a phenomenon such as schizophrenia can be defined and measured), I use as an example the 'simple' trait that we mentioned above, namely, body height. Figure 1 displays how one can measure the influence of nature and nurture on such a trait, namely, by plotting norms of reactions that ignore all kinds of causal factors except two, a genetic and an environmental one.

At issue is human body height. The displayed norms of reactions (the two grey lines) of the two genotypes (A and B) are hypothetical. Norms of reactions represent the norms of how the two genotypes at issue "behave" regarding a specific trait (e.g., body height) and given a change in environment (e.g., better nutrition over time, the period from the fifteenth to the twenty-first century).

The explanations based on such causal analysis are not mechanistic explanations in the narrow sense of explaining *how* the trait emerges. The explanations are focused on *why*, or *due to* which causes, the trait emerges, even if the developmental mechanisms are blackboxed. The explanations involve causes as difference makers.[5] Figure 1 clearly shows that nature and nurture causally interact to bring about a trait: changing the environment makes a difference to height as well as changing the genotype. Nature and nurture are both causally relevant.

Nonetheless, some theorists still want to give priority to nature. Francis Fukuyama, for instance, in *Our Posthuman Future* (2002), admits that human body height has increased over the centuries, compared to our "Middle Age" ancestors in the fifteenth century, due to improved nutrition. Fukuyama thus joins the chorus of the interactionist consensus and concedes that, yes, of course, body height is due to nature *and* nurture. A few lines later, however, he states that "the

[4]Thanks to Ken Schaffner, who provided me with this nice quote.

[5]See Woodward (2003) as one way to spell out what difference making can mean.

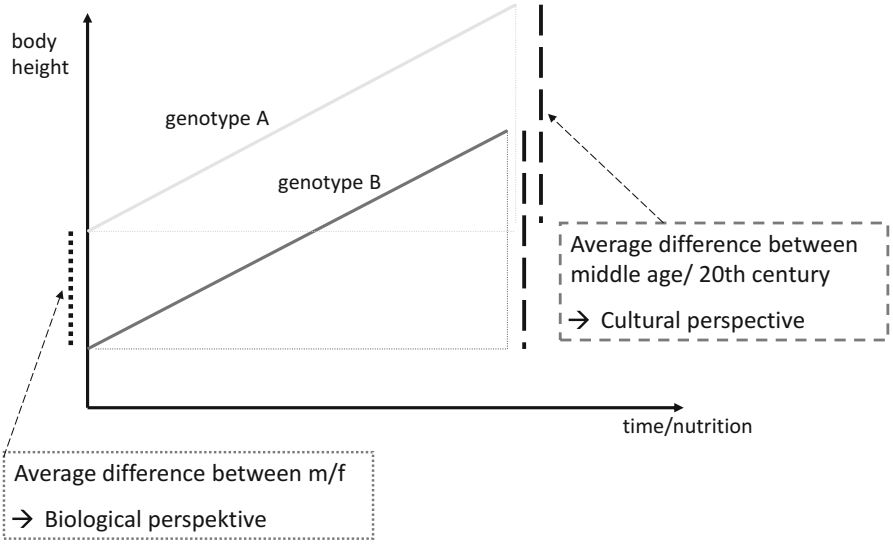


Fig. 1 Hypothetical reaction norm for human body height

average male-female differences are the products of heredity, and thus nature" (130–133). Here, he is talking about "products of heredity," i.e., he is using causal language. Is he contradicting himself? No, he is simply performing the rhetorical sleight of hand that Keller already criticized in her 2010 book, a sleight of hand that I call *implicitly changing the explanandum*, namely, a move from *explaining traits* to *explaining trait differences*.[6] What Fukuyama says is that, yes, certainly, changing the environment makes a difference to height, the *trait*, but it does *not* make a difference to the *difference between* (male and female average) heights. (The example is certainly an idealized case (see below for discussion). The hypothetical average difference between male and female body height caused by the gene is depicted by the dotted vertical line on the left-hand side of Fig. 1. Obviously, this difference is of interest to geneticists, simply on account of their disciplinary affiliation. They want to know *whether* a gene makes a difference. That's their job. They therefore look for *the difference* the gene makes a difference *to*, and ignore the rest. Others have a different job. A cultural anthropologist, for example, might reply: "Hey, that is hereditarianism and wrong, since we agreed that nutrition makes a difference too, didn't we?" – "Well, we did, but making a difference *to what*? It makes a difference to height, but it does not make a difference to *the difference between males and females*," the biologist replies, pointing to the dotted vertical line on the left-hand side of Fig. 1, which stays the same whatever you do with the depicted environmental variable. The cultural anthropologist answers: "Ah, I see,

[6]I take the distinction between traits and trait differences as explananda from Lisa Gannett (1999).

right; but wait a minute, the environment also makes a difference to a difference, but to a *different difference*: it makes a difference to what's represented by the dashed vertical line in Fig. 1 (on the right-hand side), namely, the *difference in* (average) *height* (of humans altogether) *between the Middle Ages and today*. This is the difference that interests *me* (as a historian or cultural anthropologist). It is a difference to which your genetic factor is not making any difference."

Our imaginary dialog between Fukuyama and a cultural anthropologist can also be put in terms of different *contrasts* to the explanandum. While some (e.g., Fukuyama) are interested in why men and women have (on average) different body heights (*rather* than the same), others are interested in why humans (altogether, on average) have a different height today (*compared* to 500 years ago).

The move from traits to trait differences (as explananda) amounts to an immunization against either biological or cultural perspectives. We mentioned Fukuyama as representing the one camp (being interested in genes only), and there is, for example, Alfred L. Kroeber, champion of cultural determinism, representing the other. In order to explain culture, Kroeber believed, we do not need any reference to nature, i.e., genes, etc. He did not claim that there are no biological factors causally relevant *for development* of human characteristics, but he defended *the right to ignore all biological causal factors for the cultural patterns – differences between humans in time or place – that he was interested in*.[7] There are various reasons for wishing to show the causal irrelevance of the environment (or of genes), and thereby claiming the right to ignore it. These might be reasons of academic discipline (after all, as mentioned, it is the geneticist's job to be interested in the causal relevance of genes) or political reasons (i.e., interests in maintaining the status quo and therefore downplaying the significance of certain factors in a causal explanation, e.g., poverty). Yet, in each case, the logic of the simplifying strategy stays the same.

So, in the end, we have two perspectives, a biological and a cultural one, with *different differences* as respective explananda. Those espousing these different perspectives (involving different differences and consequently different difference makers) are unable to talk to each other, since each side now has a "cause of its own" and a cause with an "effect of its own," a one-to-one monocausal explanation, despite the fact that in the world everything interacts.

The Jewish parable about the king and the archer illustrated nicely what happens when you start with a disciplinary bias toward a specific cause. Take Fukuyama again: he shoots his arrow (the cause he is interested in) and paints his explanandum accordingly. This also works the other way round: you can start with a difference you are interested in (the target) and then shoot your arrow. As the parable illustrated, the latter is more difficult. Given this, a follow-up historical study (that cannot be provided here) might show that therefore many scientific fields got

[7] See, as a very explicit and late statement on this, Kroeber (1952: 7). For details on the emergence of his cultural determinism at the beginning of the 20th century, see Kronfeldner (2009).

historically increasingly organized around preferred kinds of causal factors (and the methods suitable to study them) rather than directly around specific phenomena.

To recapitulate: scientists can adapt an explanandum (i.e., paint a target around the point to which the preferred causal factor has been shot like an arrow, using the methods available for the specific scientific field). They can do so by moving *down* to a *level of greater decomposition*, but also by moving *up* to *a level of greater abstraction*, from a trait (a property of an organism) to statistical trait differences between organisms (i.e., to the difference between averaged values of a trait constructed as a variable). This is done in order to simplify the causal complexity (many causes) at hand, i.e., in order to arrive at a partial picture that gives priority to a preferred cause of interest, even though in the world everything interacts.

4 The Pragmatic Pluralism That Results from Moving Up to a Level of Greater Abstraction

The moving-up mode of reconstituting phenomena is important in that it helps us understand how pluralism regarding causal perspectives arises, especially regarding nature-nurture debates. It also helps us understand how scientists manage to give partial explanations despite causal complexity and independently of mechanistic explanations of the developmental processes involved. Eight points are important in order to avoid misunderstanding the pragmatic-pluralistic picture resulting from this analysis.

4.1 An Idealized Case

The hypothetical norm of reaction of height that I used in the example is certainly a simple and (even more importantly) an idealized case: only if the lines run in parallel, does nurture *not* make a difference to the average difference between male and female body height. The idealization is intentionally used here, since the point I want to make is that people like Fukuyama take the idealization as a guide. In non-idealized cases, it is not sufficient to adapt the explanandum to give one's own causal factor a special, exclusive status. Yet the causal influence in which one is not interested can still be successfully downplayed, either by ignoring it completely, or by idealizing it away (assuming it to be randomized or randomizing it), or by abstracting from it (determining it as fixed). Although the "due to nature" is then less exclusive, the move from a trait to a trait difference as explanandum is still the first and necessary step toward further simplifications. If the interaction between genes and environment is nonlinear, the simplifying strategy has (so to speak) "belied" us – as Cartwright's (1983) laws of physics do – but not in an impeding manner.

It is an idealization and defines a heuristic, a research strategy, in order to arrive at some partial knowledge about the genesis of phenotypic traits. The idealized norm of reaction is intended as a simple example to illustrate the abstraction process in such cases. It shows how *two different preferred causes* or *two different choices of differences* (which are phenomena in their own right and thus explananda) place different things into the foreground or background. It thus helps to make sense of the fact that scientists ignore some causal factors, i.e., parts of the causal complexity, pertaining to the development of phenotypic traits.

4.2 It Is a Change in the Kind of Explanandum

If there is the described change in kind of explanandum, then this is more than a change from a trait token to a trait as a type. It is a change in the *kind of explanandum* (from traits to trait differences). If it were only a change from a trait token as explanandum (individual body height) to a trait type as explanandum (body height in that population), then we would simply move from explaining a token to explaining the *typical* development of the trait in a type, which would still be explaining similarity, rather than difference. In order to explain typical development of body height, you would also need *nature and nurture* interacting in order to give an adequate explanation, even in the idealized case of parallel norms of reactions. This is not necessarily the case if you explain averaged differences of height.

4.3 Difference Explanations and Trait Explanations

Even though different *kinds of explanandum* are established, the *kind of explanation* is the same, e.g., explanations involving difference makers. I take Keller (2010) to assume that with a change from traits to trait differences as explanandum, we also change the kind of explanation, because we change from causes as *trait makers* to causes as *difference makers*. Trait explanations and difference explanations would then be two categorically different kinds of explanation. Yet trait makers are simply a set of difference makers, namely, that very set that gives us (in the final instance) the *total* cause, in the sense of all the contributing causes that together are (in a given causal structure, with all its background conditions) sufficient for the to-be-explained phenomenon. If you ignore culture in a trait explanation, then you cite only a few of the many actual difference makers. If you ignore nature in a trait explanation, you do the same. What you ignore is always of the same kind: you ignore certain difference makers.

Furthermore, the question as to whether you are interested in a trait explanation or in a difference explanation depends on your pragmatic (or even political) goals: if the goal is, for instance, to *produce* something (e.g., that an enemy contracts malaria), we indeed need more than knowledge about one or only a

few difference makers. We need knowledge about the total, complete cause (a set of causes, actually). If we want to *prevent* malaria, however, we do not need such comprehensive complete knowledge of the causes. Knowledge of one of the difference makers (e.g., a mosquito bite) is enough to explain the occurrence of malaria as long as that difference maker can be regarded as a necessary condition in the context at issue (i.e., as something without which the disease would not have occurred in the context).

Thus, irrespective of whether we have a trait explanation or a difference explanation, both refer to difference makers relevant to the respective explanandum, given the pragmatic goal at hand (e.g., production, prevention, prediction, etc.). Keller is therefore right that an explanation that invokes only *one* difference maker is not the same as an explanation that gives you the total cause, a set of difference makers. But the difference is a difference in degree only – a question of how many difference makers you cite: one, two, three, many (or, all, which will usually be beyond our capacities anyway). How many you need for an adequate explanation depends on your pragmatic goals and on the differences (to which something makes a difference) into which you partition the phenomenon.

4.4 Parity at the Level of Perspectives

When biologists and cultural anthropologists, given their different interests, fight about the contribution of nature-nurture to a trait such as body height, then this can be seen as "talking at cross purposes": if they disagree about the explanandum, which they do when they settle for a division of labor along the lines of the *different differences in which they are interested*, then the explanations they respectively espouse are neither competing alternatives nor direct complements. The resulting explanations are rather alternative ways to reconstitute the original phenomenon toward different differences that do not compete with each other but can be combined later via the shared relationship of the reconstituted phenomena to the original phenomenon. There is thus a pragmatic-pluralistic parity and complementarity between the parties *at the level of the perspectives* used to study the different differences regarding a trait.

4.5 Partial, but Fruitful

If you ignore culture by changing from trait to trait differences, as Fukuyama did, you can neutralize the influence of the environmental factor by tailoring your difference. It's a trick, but it works. The knowledge generated by the different perspectives elucidates the different differences (e.g., the difference in height between the Middle Ages and now, and the alleged difference between males and females). Although the result here for each perspective is partial knowledge, it is

still a fruitful strategy to reconstitute phenomena by moving up to a level of greater abstraction. It is plausible that the division of labor we have in science, which has accelerated tremendously in the last 200 years, is itself, in part, the historical effect of the complexity of a world that we are only gradually discovering. On the way towards more of it, we can thus consider it a good thing that there are different kinds of experts devoted to their "own" kind of causes, so that there is at least some knowledge, even if it is partial (i.e., from studying one factor in isolation, while the others are ignored or held fixed or randomized).

4.6 Objectivity and Differences Worth Studying

The partial perspectives are certainly biased. Consequently, objectivity can only be reached by including them all, as partial but legitimate perspectives, as Longino (2013) illustrates in detail with respect to nature-nurture debates. Yet, in the final analysis, it all depends on whether the respective differences are worth studying since we can most likely slice up any phenomenon into as many differences as there are theoretical perspectives in science. Since there is no purely scientific way to argue that any of the differences is more important than the other, social values will have an influence on that. For instance, as Kourany (forthcoming) stresses regarding cognitive difference research, a society might well decide that certain differences (e.g. between races or gender) should not be studied anymore, for social reasons.

4.7 Integration at the Level of Traits

The integration of the partial difference explanations on the basis of the described division of labor most certainly occurs at the level of the trait that has been fragmented into all kinds of *differences regarding the trait*. In the long run, a division of labor and integration of the partial causal knowledge are both required for the increase of our knowledge about phenotypic traits. It takes two to tango.

4.8 But Can the Different Difference Explanations Be Integrated?

A radical pluralism, such as advocated by Longino (2013), denying the possibility of integration of the diverging perspectives in the nature-nurture debates, is inadequate. It is a pragmatic, integrative pluralism that is advocated here. Integration of the causal knowledge arrived at by looking at causal factors in isolation is possible, as any norm of reaction exemplifies. We can combine the knowledge *that* a

specific genotype makes a difference with regard to a difference in body height (in the measured situation) and the knowledge *that* nutrition influences a *different difference* regarding body height (with the same limitations), even in non-idealized, i.e., non-linear cases. We might even learn something about further patterns of dependence, as illustrated in the gene-interaction cases studied by Caspi et al. (e.g., 2002) or as described by Kitcher (2001).

Longino (2013), however, demands that true integration be not only an integration of knowledge *that* the respective factors make a difference but also an integration of knowledge about *how much* difference these make. She assumes that approaches involved in nature-nurture debates always want an answer to a "how much" question, which I think is a mischaracterization. In my opinion, she mistakenly transfers the 'how much' approach, as used in quantitative behavioral genetics, to other disciplines for the study of humans. Such a quantitative integration is indeed impossible, as the now classic discussion on apportioning causality with respect to nature-nurture shows. [8] Yet few scientists, I would say, care for it as a goal (except certainly traditional quantitative behavioral geneticists). Furthermore, true integration for Longino is only happening if *all* the partial knowledge derived from different perspectives is combined into a "single comprehensive account" (Longino H, December 12, 2014, personal communication).[9] This sets the standards so high so that integration is – simply because of that standard of complete knowledge – usually impossible since it is usually beyond human capacities to give complete causal accounts. Finally, for the social aims at issue in nature-nurture debates, "how much" questions and a completeness standard are not important either, given that interventionist social policies can operate with less than "how much" knowledge integrated into a complete account of the phenogenesis of the trait.

From my point of view, Longino's picture demands too much and – as a result – is too pessimistic, as if there were no way to deal with uncertainty regarding "how much" nature and nurture contribute to a trait. In the final instance, integrating difference explanations should elucidate *how* a trait such as body height is produced mechanistically (i.e., stepwise), without being able to say *how much* each factor contributes in relation to each other, i.e., comparatively, and however partial that mechanistic explanation is. In sum, the above-mentioned nonquantitative partial integration – of knowledge *that* certain factors make a difference (to different differences) regarding a trait – can contribute to mechanistic explanation, even though the difference explanations are themselves not necessarily mechanistic.[10]

---

[8] J. St. Mill (1858) already acknowledged the problem under the label of a "composition of causes." See, for instance, Sober (1994), Keller (2010), and Walsh (2013) on this classic problem of apportioning "how much" in nature-nurture debates.

[9] Compare Tabery et al. (2014), where Longino stresses in reply to Tabery that anything short of a "complete and comprehensive" account of the mechanisms involved in the phenogenesis of a trait is not adequate.

[10] See Tabery (2009) for more on integrating mechanistic explanation and the search for differences. See also the exchange in Tabery et al. (2014).

My perspective on integration is thus similar to Sandra Mitchell's (2003) integrative pluralism. However, since Mitchell is not concerned with nature-nurture issues, she seems to ignore that there are cases of partial causal perspectives where we can partition the explanandum and divide it so that it can be *shared* (being related to the same trait), without having the *exact same explanandum* (by being focused on *different differences*). In her integrative pluralism (as in Longino's nonintegrative pluralism), there seem to be different questions and different causes (derived from incongruent different perspectives), but only one trait and no trait differences mediating the integration of the different causes.

Keller (2010), by contrast, falls short in the exact opposite direction, by situating difference explanations and trait explanations too far apart, albeit while acknowledging the distinction between a *trait* and *trait differences* as explanandum.

5 Summary

This paper addressed the epistemology of simplifying explanatory strategies. It focused on whether we can make sense of the fact that scientists ignore some causal factors involved in the explanation of a phenomenon by *reconstituting their explanandum phenomenon* (i.e., by redefining it so that it better fits the aims with which one started). The answer is that, as a pragmatic-pluralist, we can. This paper also explicated how disciplinary boundary politics enter the business of reconstituting phenomena: disciplinary perspectives constitute the "arrows" (i.e., the causal factors at reach or preferred from that perspective) that are "shot" and according to which general and shared explananda (i.e., complex traits such as body height, schizophrenia, etc.) are adapted so that they can be "bull's-eyes" (i.e., explananda that are less causally complex and fitting to the arrows shot). Depending on case, the reconstitution of phenomena will be down a level of composition (e.g., in the case of the concept of endophenotypes) or up a level of abstraction (e.g., in the case of explaining different differences).

Acknowledgements I want to thank the *Center for Philosophy of Science* at the *University of Pittsburgh* and Martin Carrier for their great support during the time this paper was written. I also want to thank Bill Bechtel, Uljana Feest, Lisa Gannett, Peter Godfrey-Smith, Jens Harbecke, Evelyn Fox Keller, Helen Longino, Sandra Mitchell, Ken Schaffner and one of the two anonymous referees for interesting and helpful feedback. I want to particularly thank Alexander Reutlinger for the many inspiring discussions related to the topics of this paper.

References

Bechtel, W., & Richardson, R. C. (1993/2000). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton: Princeton University Press.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97*, 303–352.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon.
Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., et al. (2002). Role of genotype in the cycle of violence in maltreated children. *Science, 297*, 851–854.
Feest, U. (2011). What exactly is stabilized when phenomena are stabilized? *Synthese, 182*, 57–71.
Fukuyama, F. (2002). *Our posthuman future*. New York: Farrar, Straus and Giroux.
Gannett, L. (1999). What's in a cause?: The pragmatic dimensions of genetic explanations. *Biology and Philosophy, 14*, 349–373.
Goldstein, D. B. (2009). Common genetic variation and human traits. *New England Journal of Medicine, 360*, 1696–1968.
Gottesman, I. I., & Gould, T. D. (2003). The endophenotype concept in psychiatry: Etymology and strategic intentions. *The American Journal of Psychiatry, 160*, 636–645.
Keller, E. F. (2010). *The mirage of a space between nature and nurture*. Durham: Duke University Press.
Kitcher, P. (2001). Battling the undead: How and (how not) to resist genetic determinism. In R. Singh, C. Krimbas, D. Paul, & J. Beatty (Eds.), *Thinking about evolution: Historical, philosophical and political perspectives* (Vol. 2, pp. 396–414). Cambridge: Cambridge University Press.
Kourany, J. (forthcoming February 21). *Should some knowledge be forbidden? The case of cognitive differences research* [Draft]. Presented at the Fishbein Workshop in the History of the Human Sciences, Chicago.
Kroeber, A. L. (1952). *The nature of culture*. Chicago: The University of Chicago Press.
Kronfeldner, M. (2009). If there is nothing beyond the organic . . . : Heredity and culture at the boundaries of anthropology in the work of Alfred L. Kroeber. *NTM – Journal of the History of Science, Technology and Medicine, 17*, 107–133.
Longino, H. E. (2013). *Studying human behavior : How scientists investigate aggression and sexuality*. Chicago: The University of Chicago Press.
Mill, J. S. (1858). *A system of logic, ratiocinative and inductive*. New York: Harper & Bros.
Mitchell, S. D. (2003). *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.
Schwartz, S. (1998). The role of values in the nature/nurture debate about psychiatric disorders. *Social Psychiatry and Psychiatric Epidemiology, 33*, 356–362.
Sober, E. (1994). Apportioning causal responsibility. In *From a biological point of view: Essays in evolutionary philosophy* (pp. 184–200). Cambridge: Cambridge University Press.
Tabery, J. (2009). Difference mechanisms: Explaining variation with mechanisms. *Biology and Philosophy, 24*, 645–664.
Tabery, J., Preda, A., & Longino, H. (2014). Pluralism, social action and the causal space of human behavior. *Metascience, 23*, 443–459.
Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics, 40*, 489–490.
Walsh, D. M. (2013). The negotiated organism: Inheritance, development, and the method of difference. *Biological Journal of the Linnean Society, 112*, 295–305.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Manipulating Spins: Causality and Decoherence

Fernanda Samaniego

1 Introduction

In the so-called "spin-echo experiments" a set of spins is affected by several magnetic changes. The resulting behavior of the spin's system seems to violate the second law of thermodynamics. For this reason the spin-echo experiments are considered of particular interest for the foundations of physics and several explanations of them have been put forward. Hemmo and Shenker (2005) offer a causal explanation, according to which, the approach to equilibrium is ultimately produced by the environmental perturbations acting upon the system trough decoherence.

The decoherence-based interventionist explanation was assessed using Woodward's (2003) manipulability theory of causal explanation. The analysis (in Samaniego 2013) revealed that the correlation between the two causes postulated by Hemmo & Shenker seemed to render their explanation 'shallow' –in accordance with the criteria of the manipulability theory– in the sense that no prove could be provided for the postulated causal relationship being genuine.

Fortunately, in a recent paper (2011) Woodward himself has offered a strategy to resolve a similar case with non-independent causes. Basically, Woodwrad suggests to design *joint interventions* for performing simultaneous interventions upon the two correlated casual variables.[1]

[1]Woodward proposes two different strategies in his paper (2011). We are referring to the second one.

F. Samaniego (✉)
Department of Philosophy and Literature, National Autonomous University of Mexico, Mexico City, Mexico
e-mail: Fernanda.Samaniego@gmail.com

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_14

The main objective of this paper is applying such a strategy to our case of study of the spin-echo. By doing so we will find interesting outcomes. One, the present analysis indicates that a manipulability theory of causation may offer promise as a mean of assessing causal patterns in the social and health sciences where the Causal Markov Condition is not fulfilled. Additionally, if the causal relationship between decoherence and spin-echo is genuine, we have good reasons to explore the possibility of using decoherence as a fundamental explanation of entropy rising.

2 Manipulability Definition of Cause, Intervention and Invariance.

The Manipulability theory (Woodward 2003) is both a theory of causation and a model of causal explanation. It provides criteria with which to detect when a given causal relationship is genuine and, concomitantly, when a given explanation –based on that causal relationship– counts as satisfactory. The objective in this section is to present the manipulability understanding of *genuine causal relationship*. In order to do so, the manipulability notions of *cause*, *intervention* and *invariance* must be defined; these constitute he central notions of the manipulability theory that will be used in the subsequent sections.

Cause The manipulability notion of cause is defined as follows: "C is a genuine cause of E if, given the appropriate background conditions, there is a possible manipulation of the cause C such that this is also a way of manipulating or changing the effect E" (see Woodward 2003, section 2.2; or Woodward 2008, section 1). In other words, causal relationships entail some changes upon the values of E whenever the values of C are modified. According to the theory, the manipulations carried out over C must be reproducible in the sense that responses to the effect E must be in some way repetitive or systematic.

Intervention Interventions, in turn, are understood in the manipulability theory as "exogenous causal processes that change the value of the putative cause C in such a way that if any change occurs in the effect E, it occurs only in virtue of E's relationship to C and not in any other way" (see Woodward 2003, p. 47). The manipulability notion of *intervention* is formally defined as follows:

(IN) I assuming some value $I = i$ is an intervention on C with respect to E if and only if $I = i$ is an actual cause of the value taken by C, and I meets the following conditions:

(IN-i) I must be the only cause of C; i.e., the intervention must completely disrupt the causal relationship between C and its previous causes so that the value of C is set entirely by I.

(IN-ii) I should not itself be caused by any cause that affects E via a route that does not go through C.

(IN-iii) *I* must not directly cause *E* via a route that does not go through *C*.
(IN-iv) *I* leaves the values taken by any causes of *E* except those that are on the directed path from *I* to *C* to *E* (should this exist) unchanged (Woodward 2010, section 5; and 2003, p. 98).

Interventions are taken as performable *in principle* or hypothetically. They seek to capture the ideal experimental conditions that should be fulfilled to change the value of *C* and so study its causal link with *E*. The interventions that we use to test a given causal relationship are also referred in the manipulability theory as *testing interventions.*

Invariance Finally, the manipulability notion of intervention is defined as follows: Let *G* be a generalization relating changes in the cause *C* (from *c* to c_*) to changes in the effect *E* (from *e* to e_*). We say that *G* is invariant under a testing intervention if and only if it correctly describes what the new value of *E,* e_*, would be under this change; that is, if and only if it remains true that $G(c_*) = e_*$ (see Woodward 2003, section 6.2).

Genuine Causal Relationship Using the three notions defined above, manipulability theory specifies that, the causal relationship between *C* and *E* is *genuine* if it remains invariant under a set of interventions (see Woodward 2003, section 1.4).

In other words, performing interventions allows us to determine whether or not *C* causes *E.* If for a given claim "*C* causes *E*" there are no well-defined interventions that would change the value of *C*, then we have no means to prove, via the manipulability theory, that the relationship between *C* and *E* is genuinely causal. This last fact will be central for the ideas developed in the rest of the paper.

3 Problematic Correlated Causes

3.1 Correlated Causes in the Spin-Echo Case

Let me briefly describe the spin-echo experiments.[2] In a spin-echo experiment a set of nuclear spins, normally belonging to protons in a sample of glycerin, are placed in a strong magnetic field. Through the application of a first radio-frequency pulse, the spins are initially aligned. Let us say, for example, that the direction of the magnetic field is in the z-axis while the superposed spins lie over the xy-plane and they are all pointing in the same direction. Due to the presence of the magnetic field all the spins are initially precessing with the same frequency and this produces the emission of an electromagnetic signal.

The spins are then left to evolve for a while and the discontinuities in the magnetic field cause slight differences in the precession rates of the spins. The

[2]This is a summary, for more details see Hemmo and Sheker (2005) and Samaniego (2013).

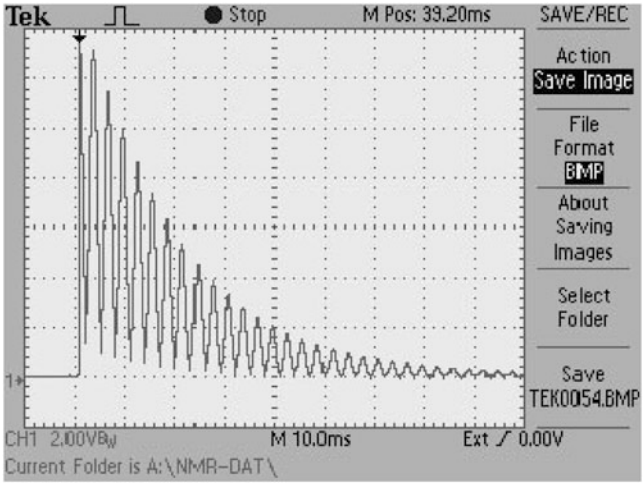


Fig. 1 Spin-echo decay

precession rate of each spin is more or less affected depending on the strength of the field at each point. After an interval of time (τ), the spins reach a disordered state, in which they are not pointing in the same direction anymore and the electromagnetic macroscopic signal completely disappears. In a second part of the experiment a reversal is induced by another radio-frequency pulse. After an interval of time (2τ) the spins are realigned and this causes the re-emission of the electromagnetic signal. In that moment the intensity of the electromagnetic signal reaches again a maximum point. The repetition of the signal is the phenomenon that gives the name to the experiments.

As shown in Fig. 1, the height of the echo signals is not constant. In fact, it always decreases. And, after an interval of time (known as relaxation time), the system is unable to generate the echo-signal.

Spin-Echo experiments are considered as a particular challenge in foundations of statistical mechanics. The reason is that,, during the experiments, entropy seems to increase and decrease repeatedly. Several philosophers of physics provided explanations of the spins behavior in these experiments (Blatt 1959; Ridderbos and Redhead 1998; Hemmo and Shenker 2005). However, in this paper we focus in Hemmo and Shenker's answers to the two following questions: How can we explain the decay of the echo signal? In other words, what prevents the return of the spins to their original state?

According to the decoherence-based explanation of the spin-echo experiments provided by these authors (2005), the perturbations upon the spins' system constitute the very cause of the decay of the echo signal. The decoherence-based approach combines two kinds of perturbations: the external decoherence associated with the environment, and the internal decoherence associated with the spin-spin interaction.

External decoherence (*DE)* emerges from the interaction between the spins and their environment, where by "environment" we mean the magnetic field, where the glycerin sample is immerse. As described by standard decoherence models, the pointer basis of *DE* is *position* and the states of the environment relative to different system positions of the system become approximately orthogonal. This means that glycerin protons' spin state is decohered through its dependence on position (for more details see Hemmo and Shenker, 2005, p 641).

Internal decoherence, (*DI*), on the other hand, emerges from spin-spin interactions which directly affect protons' spin state. Here the 'pointer basis' is *spin*, and this time the "environment" of each spin is understood as the rest of the spins in the sample (for more details see Hemmo and Shenker 2005, p 641).

Let us express and analyze the decoherence-based explanation in terms of Woodward's manipulability theory.[3] The set of relevant variables can be defined as *DE, DI, EC, H,* where *DE* = rate of external decoherence; *DI* = rate of internal decoherence; *EC* = effective collapses; *H* = height of the echo-signal. And the causal relationships are given by the directed graph in Fig. 2.

According to the manipulability theory, the values of all the variables causally connected with the effect *H* must be fixed. This means that every time we vary the value of the *DE* the value of the internal decoherence *DI* should remain fixed. This is however impossible. Due to the coupling between the spin and the position, the internal and the external decoherence are not independent from each other. Manipulating the magnetic field (external decoherence) automatically provokes a change in the directions of the spins (internal decoherence). In the same way, manipulating the spin-spin interaction, will necessarily affect the magnetic field around them, and therefore the total magnetic field of the sample, and its relation with the external environment, is affected as well. This means that it is impossible to block one kind of decoherence in order to effectively manipulate the other one.

The conclusion of our causal analysis of this case is the following: We cannot prove that the causal relationship between decoherence and the echo-decay is genuine, for we cannot provide interventions under which such causal relationship remains invariant.

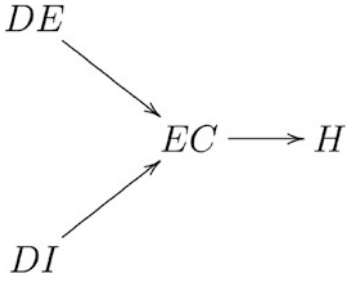


Fig. 2 Causal graph portraying the decoherence-based explanation of the decay and disappearance of the echo-signal. *Arrows* indicate causal relationships

[3]This is a summary, full analysis in Samaniego (2013).

3.2 Correlated Causes in the Cholesterol Sase[4]

Let us now turn to an assessment of the causal relationship between cholesterol and arteriosclerotic vascular disease (*ASVD*). Among all the sterols in a human body, two are particularly relevant factors as regard heart disease: high-density lipoproteins and low-density lipoproteins. The low-density lipoproteins (*LDL*) are composed by small amounts of proteins and large lipid quantities. If the organism does not properly eliminate *LDLs*, they accumulate on blood vessel walls, forming atheromas. These atheromas may reduce or even obstruct blood flow, preventing the heart and brain's proper oxygenation. Additionally, oxidized *LDL* molecules damage and weaken the vessel walls. These are the reasons why *LDL* are often referred as "bad cholesterol". "High-Density Lipoprotein" (*HDL*) molecules, on the contrary, are highly protein dense, and low in lipids, and thus *prevent* heart diseases. Most of the *HDL* in our bodies are produced in the liver and they are transformed in bile for digenstion. *HDL* also circulates in the blood, taking on cholesterol that cells have not use. hey travel to the liver and are broken down and recycled for bile production. For such reasons *HDL* is also known as "good cholesterol."

Using this information, one could propose the causal graph in Fig. 3 as a portrayal of the causal relationships between good cholesterol (*HDC),* bad cholesterol *(LDC)*, and atherosclerosis (*ASVD).* Then we verify if our proposed interventions meet the conditions **IN**. And we find a difficulty analogous to the difficulty in the spin-echo case: According to requirement **IN-iv**, *HDL* values should remain fixed while we tweak *LDL* values, and vice versa. However, due to the intertwined relationship between *HDL* and *LDL,* this seems impossible; *HDL* and *LDL* levels may influence each other. For example, high amounts of *LDL* in the blood may inhibit the production of *HDL* in the liver. Or, in a similar yet inverse fasion, if *LDL* levels break down drastically, the organism may perceive lack of cholesterol in the blood and an increase in *HDL* production. This interrelation between *HDL* and *LDL* levels

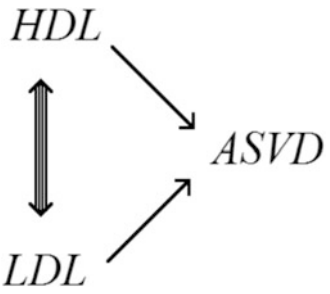


Fig. 3 Causal graph relating the two types of lipoproteins and arteriosclerotic vascular disease. *Arrows* indicate causal relationships. *Double arrow* indicates causal co-dependence

[4]The correlation in the cholesterol case was originally pointed out in Sprites and Scheines 2004. Sprintes and Scheines present the problem from a different perspective. However, we all agree in the fact that defining interventions in this case is highly problematic.

seems to prevent all the dietetic manipulations to meet requirement **IN-iv.** The double arrow connecting *HDL* and *LDL* in Fig. 3 represents the interdependence between those two variables.

In violation of requirements **IN,** the proposed interventions of cholesterol based on diet modifications do not classify as proper interventions.

Therefore, in analogy to the conclusion in the spin-echo case, our causal analysis lead us to conclude the following: As far as we have no interventions under which the causal relationship remains invariant, we have no means to prove that the causal relationship between cholesterol and the arteriosclerotic vascular disease is genuine (or in any case, the manipulability theory has not provide us with those means).

3.3 Causal Markov Condition

Last but not least, it is important to note that both the spin-echo case and the cholesterol case fail to meet the *Causal Markov Condition,* specifically, that two causal parents are not independent.[5]

4 Joint Interventions: Woodward's Strategy to Solve the Cholesterol Case

At this point one could simply assume that Woodward's interventionism is not applicable to cases with multiple correlated causes. One could thus dismiss Woodward's theory and look for alternative accounts of causation. That said, one may also choose to leave the manipulability theory in place; my motivation to follow this latter alternative is Woodward's flexible attitude towards possible changes and improvements in his own theory (see, for example, 2003, p. 132, and 2011). This attitude has provided me a strong incentive to propose here an attempt at applying the manipulability theory to cases where it first seemed problematic to do it.

The challenge that emerges from our cases study, namely, form the spin-echo case, is whether the manipulability theory can tell us something about the causal link between a given effect E and its putative causes C_1, C_2, C_n, when those causes are correlated to each other. Even in those cases, Woodward believes that we can legitimately use counterfactuals to elucidate causal claims along the lines that the manipulability theory suggests.

[5]I'm following here the definitions of Causal Markov Condition in Suárez and San Pedro (2010, p. 9); and Hausman and Woodward (1999, p. 523).

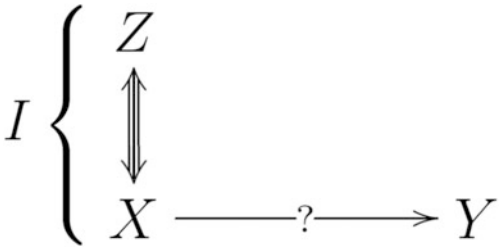


Fig. 4 The simple *arrow* represents the putative causal relation, *while* the *double arrow* stands for the joint intervention acting upon variables *Z* and *X*. Therefore, *Z* and *X* are not independent from each other (Figure taken from Woodward, 2011)

In such cases, Woodward suggests verifying whether two conditions –specifically the two original motivations for introducing the notion of intervention– are met. First, we must ensure that "there is a basis for claims about what will happen to *E* under an intervention on *C*"; i.e. we should be able to associate some well-defined notion of change with *C*, and we have some grounds for saying what the effect, if any, on *E* would be of changing *C* only – and nothing else. Secondly, there must be "a way of disentangling the effect on E of changing just *C* from the effects on *E* of changes in other potentially confounding variables" (see Woodward 2003, p. 131–132). In Woodward's view, all these two considerations will give us a purchase on whether the counterfactual "if *C* were to be changed by an intervention *I*, then *E* would change" is true or false (see Woodward 2003, p. 132).

In a more recent paper (2011), Woodward also shows his willingness to modify and improve the manipulability theory. He assesses the cholesterol case and suggests understanding a single intervention that acts simultaneously in the two co-dependent variables *HLD* and *LDL*. He proposes to identify *joint interventions* upon these two variables as shown in Fig. 4. (2011, p. 21).

As Woorward points out, "In some cases a combination of interventions on several variables, although co-possible, may force [. . .] a change in additional variables as well. We should either (a) disallow such combinations of interventions for the purposes of assessing causal relationships or (b) somehow compensate for their effects in tracking causal relationships, taking care to avoid double-counting" (2011, p. 21).

5 Applying the Same Strategy to the Spin-Echo Case

In this section I will follow Woodward's three rules in order to develop a new test of the putative causal relationship between two types of decoherence and spin-echo decay (namely, the causal relationship represented in Fig. 2, in Sect. 2).

First of all, in accordance with Woodward's strategy, we must identify the interdependence between internal decoherence *DI* and external decoherence *DE*. As mentioned before, *DE* is associated to the quantum observable position, while *DI* is associated to the quantum observable spin. In the spin-echo experiments, therefore, the relationship between variables *DE* and *DI* is the outcome of the quantum correlation between the observables spin and position.

Secondly, we must propose joint interventions that can act simultaneously on variables *DE* and *DI*. To do so we must design possible ways of reducing and increasing decoherence levels. Fortunately for us, many researchers in quantum information have already been working on designing ways to control decoherence levels. Reducing decoherence enables to retain quantum information; and completely avoiding decoherence would be ideal for building the so-called quantum computer (see Uhring 2009).

For these reasons, suppressing decoherence in spin-echo experiments has lately become a research field of great interest. And during the last few years, several different sequences of radio-frequency pulses have been put forward seeking to reduce decoherence as much as possible or for as long as possible. This situation has been experimentally developed with spins obtaining successful results. In the latest models, the optimized sequence of pulses in spin-echo experiments, is said to efficiently suppress decoherence getting important prolongations of the relaxation times (see Uhring 2009 and references there in).

For the purposes of our causal analysis it would also be very convenient to find a joint intervention, upon the variables DE and DI, that represses both types of decoherences throughout the entire experiment. In such a situation, according to Hemmo and Shenker, the spin system would never approach equilibrium. One way of preventing both decoherences is to set up the system in a "non-decoherent quantum state." Such a state has not yet been physically performed and is considered an experimental impossibility at present. It is nonetheless a theoretically conceivable quantum state, and is valid as a hypothetical joint intervention in the present analysis.

A much simpler and currently more performable intervention would be to dissolve the glycerin sample in water. If we add water to the sample, the distance between the glycerin molecules increases and each spin moves away from the others. Hence, if the decoherence approach is correct, both internal and external decoherences would decrease, slowing down the system's approach to equilibrium.

An experiment comparing the Spin-Echo signal for different dilutions of water and glycerin has been recently performed (Hughes 2005). The results show that the relaxation time of the spins' system actually depends on the water-glycerine proportion in the sample: the higher the glycerin content, the shorter the relaxation time. In other words, the experimental results are compatible with Hemmo and Shenker's decoherence-based explanation.

In sum, we have proposed the following joint interventions:

I1 = Reducing decoherence through multi-pulse sequences;
I2 = Preparing a "non-decoherent" quantum state (hypothetical manipulation);
I3 = Reducing glycerine sample viscosity.

Finally, in accordance with Rule 3, we need to specify how double counting will be avoided. In this case it requires both defining the system's density matrix, and understanding how the position-spin quantum correlation acts. The glycerine sample properties, the external field, and spin-spin interactions will determine the density matrix specifics. Hemmo and Shenker (2005) sketch how this density matrix should be.

6 Final Remarks

Internal and external decoherences are so intricately related that every attempt to manipulate one of them will necessarily affect the other. However, in analogy with the "cholesterol case" assessed by Woodward (in 2011), we discovered that understanding the causal structure allows us to propose joint interventions to perform a manipulability test of the relationship between decoherence and Spin-Echo decay.

Two main consequences emerge from this paper's analysis:

First, that manipulability theory is being applied to cases where the Causal Markov Condition is not met. If applicable, such cases may represent an important advantage of the manipulability theory over other approaches to causation, since causal explanations in the natural, social and heath sciences often fail to meet the Causal Markov Condition.

Second, this paper goes on to prove that decoherence and Spin-Echo decay are genuinely causally related (in the manipulability-related sense of a genuine causal relationship). Dependence between the two types of decoherence has been clearly defined. Joint interventions have been properly defined and experimental evidence has been provided for two such interventions. And, to complete the proof, Woodward suggests avoiding double counting. This requires formal work on density matrices that Hemmo and Shenker (2005) have already sketched out. There are physicists both in New Zealand and Mexico who have expressed interest in helping to undertake this task. Proving that the causal relationship between decoherence and the spin-echo decay is genuine, is of particular interest for Foundations of Physics, since it provides support for a foundational explanation of entropy behavior based on quantum features.

Acknowledgements I am deeply grateful to Mauricio Suárez, Bert Leuridan, Federica Russo, Phyllis Illari, Ana Rosa Pérez-Ransanz and Elías Okon for their kind support and bibliographical recommendations.

References

Blatt, J. M. (1959). An alternative approach to the ergodic problem. *Progress of Theoretical Physics, 22*(6), 745–756.

Hausman, D., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science, 50*(4), 521–583.

Hemmo, M., & Shenker, O. (2005). Quantum decoherence and the approach to equilibrium II. *Studies in History and Philosophy of Modern Physics, 36*, 626–648.

Hughes, P. (2005). *Spin echo nuclear magnetic resonance* (Laboratory report). Department of Physics and Astronomy, The University of Manchester. http://porlhews.tripod.com/sitebuildercontent/sitebuilderfiles/Spin_Echo_NMR_lab.pdf

Ridderbos, T. M., & Redhead, M. L. G. (1998). The spin-echo experiments and the second law of thermodynamics. *Foundations of Physics, 28*(8), 1237–1270.

Samaniego, F. (2013). Causality and intervention in the spin-echo experiments. *Theoria, 21*(3), 477.

Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science, 71*(5), 833–845.

Suárez, M., & San Pedrio, I. (2010). EPR, robustness and the causal Markov condition. In M. Suárez (Ed.), *Causes, probabilities and propensities in physics* (Synthese library, Vol. 347, pp. 173–193). New York: Springer.

Uhring, G. (2009). Concatenated control sequences based on optimized dynamic decoupling. *Physical Review Letters, 102*, 120502.1–120502.4.

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

Woodward, J. (2008). Causation and manipulability. Stanford encyclopaedia of philosophy. http://plato.stanford.edu/entries/causation-mani/

Woodward, J. (2010). Scientific explanation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2010 Edition). http://plato.stanford.edu/archives/spr2010/entries/scientific-explanation/

Woodward, J. (2011). Interventionism and causal exclusion. http://philsci-archive.pitt.edu/id/eprint/8651

Part V
Philosophy of Physics and Chemistry

How Fundamental Physics Represents Causality

Andreas Bartels and Daniel Wohlfarth

1 Introduction

Russell's dictum that there is no place for causality in fundamental physics (Russell 1912/1913) has been revitalized in a recent debate (see, e.g., the contributions in (Price and Corry 2007)). This debate combines a rather heterogeneous collection of approaches to causality coinciding essentially in one common focal point: their approval of Russell's thesis.

Russell's main reason for denying a genuine place for causality in physics was that the *asymmetry* of the causal relation – if A causes B, then it is not the case that B causes A – have no counterpart in modern theories of physics. The laws figuring in those theories make no difference whatsoever concerning the determination of the state of a system by either its past or its future states. After having declared that it is not the "same" cause producing the "same" effect, but "same" (invariant) *relations* given by differential equations that represent the so-called "law of causality" in physics, Russell claims concerning this law:

> The law makes no difference between past and future: the future "determines" the past in exactly the same sense in which the past "determines" the future. The word "determine", here, has a purely logical significance: a certain number of variables "determine" another variable if that other variable is a function of them. (Russell 1912/1913, p. 15)

A first and longer version of this contribution has been published in: Maria Carla Galavotti et al. (eds.): *New Directions in the Philosophy of Science*, Springer 2014.

A. Bartels (✉) • D. Wohlfarth
Institute for Philosophy, University of Bonn, Regina-Pacis-Weg 3, 53113 Bonn, Germany
e-mail: andreas.bartels@uni-bonn.de; nullgeodaete@aol.com

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_15

Since determination by laws of physics is always a symmetric relation, there is no hope that we can find a basis for causal asymmetries in those laws. This argument is the starting point for Huw Price's (2007) approach towards causality. The basis of causal asymmetry cannot be found within the laws of physics. Nor can the alleged de facto asymmetries between initial and future conditions produce any acceptable physical alternative. Since de facto asymmetries do not represent a fundamental trait of the universe, but reflect only the particular thermodynamic conditions which dominate the behavior of macroscopic systems in the low entropy era that we observe now, they cannot provide a physical basis of causal asymmetry either. Thus the basis of causal asymmetries has to lie outside physics. According to Price, it is constituted by the structure of human agency.

We will argue that despite the premise used by Russell in his argument against the fundamentality of causality is essentially correct, the conclusion does not follow: causality, contrary to Russell's claim, has a basis in fundamental physics.

It has to be noticed that Russell attacks only one particular way in which causality could be anchored in fundamental physics, namely the way of being represented by fundamental equations. This sort of anchoring, Russell claims, is forbidden by the symmetry of determination relations as provided by fundamental equations. But there exists a further way of fundamental anchoring of causality. As we shall argue in Sect. 2, despite the time-reversal invariance of fundamental laws it is possible that the solutions of those laws are *typically*[1] *time-asymmetric*.[2] In fact, it has been proven that *almost all* spacetimes that are solutions of the field equations of General Relativity and which allow for a "cosmic" time parameter – that is, spacetime seperates into curved space and a universal time common to all comoving observers – and, furthermore, possess a matter field are time-asymmetric.[3]

That a spacetime possessing a cosmic time (and therefore having spacelike hypersurfaces) is *time-symmetric* with respect to a hypersurface $t = t_S$, intuitively means that, from the hypersurface $t = t_S$, the "spacetime looks the same in both temporal directions. Therefore, if a time-orientable spacetime having cosmic time is time-asymmetric, we shall not find a spacelike hypersurface $t = t_S$ which splits the spacetime in two 'halves', one the temporal image of the other one with respect to their intrinsic geometrical properties."[4] On the other hand, if we find such a hypersurface, we shall call the spacetime time-symmetric.

[1] In Sect. 3, we will elaborate on the notion of typicality in connection with the asymmetric behavior of solutions of the field equations of GR.

[2] cf. (Castagnino and Lombardi 2009, p. 3).

[3] cf. (Castagnino et al. 2003, p. 900 f.), (Wohlfarth 2013).

[4] cf. (Castagnino and Lombardi 2009, p. 14.) In more technical terms, time-symmetry of a spacetime with respect to a spacelike hypersurface $t = t_S$ requires "time-isotropy", i.e. the existence of a diffeomorphism onto itself which reverses the temporal orientations but preserves the metric and leaves the hypersurface $t = t_S$ fixed.

In Sect. 3, we will show that the result, according to which *almost all* spacetimes that are solutions of the field equations of General Relativity and which allow for a universal cosmic time parameter and, furthermore, possess a matter field are time-asymmetric, provides a new resource for anchoring the causal asymmetry in physics and thus diminishes the need for epistemic[5] or even anthropocentric foundations of causality.

2 The Time-Asymmetry of General Relativistic Spacetimes

The first step, in this Section, is to argue that *almost all* spacetime models of general relativity are time-asymmetric (cf. Castagnino et al. 2003; Wohlfarth 2013). Thereby we agree with the general proposal in Maudlin's "On the Passing of Time" (Maudlin 2007a), according to which "[T]he passage of time is an intrinsic asymmetry in the temporal structure of the world" (Maudlin 2007a, p. 108), and also to Earman's *Time Direction Heresy* entailing the claim that "a temporal orientation is an intrinsic feature of space-time which does not need to be and cannot be reduced to nontemporal features" (Earman 1974, p. 20). In particular, like Maudlin we deny any attempt to reduce the direction of time to the increase of entropy.[6] But we differ with respect to how the "intrinsic asymmetry in the temporal structure of the world" is established and what it is.

Maudlin argues that the entropic atypicality[7] of microstates with respect to their backward temporal evolution (temporal evolution in backward time leads to lower entropy), in contrast to the typicality of their behavior in forward time direction, can only be explained by means of their causal production: "The atypical final state is accounted for as the product of an evolution from a generically characterized initial state." (Maudlin 2007a, p. 133) But since "[T]his sort of explanation requires that there be a fact about which states produce which" (Maudlin 2007a, p. 134) there must be an intrinsic direction of time providing the needed causal asymmetry.

However we judge the explanatory value of Maudlin's solution for the problem of the direction of time, the very start of his argument takes for granted that the world is asymmetric in the sense that the global entropic behavior shows up typicality with respect to one temporal direction, but atypicality with respect to the other temporal

[5]Frisch has provided an impressive epistemic account of causality in fundamental physics (cf. Frisch 2012).

[6]Castagnino et al. have raised an objection against this reductionist move: The entropy of the universe can only be defined under some physical conditions referring to space-time as a whole, in particular the condition that the space-time allows for a foliation into space-like hyper-surfaces, e.g. there exists a cosmic time (cf. Castagnino et al. 2003, p. 896).

[7]See for the notion of typicality: (Maudlin 2007b) and (Goldstein 2012). Our own proposal, provided in this Section, will also make reference to this notion.

direction. Now, as shown by Castagnino et al. (2003), *global* entropy can only be defined in universes with a cosmic time. This seems to indicate that in establishing an intrinsic time-asymmetry of the universe, cosmological accounts have conceptual priority over entropic accounts. Thus, we choose a cosmological account in order to establish intrinsic time-asymmetry.

We differ from Maudlin's account also concerning what this intrinsic asymmetry *is*. As far as we can see, there is no definite answer to this question in Maudlin's work. He claims that "all that seems to be required ... is an orientation." (Maudlin 2007a, p. 135) But it remains unclear, at least to us, what intrinsic structures of spacetime, according to his account, actually yield such orientation.

Now, we start our account by showing that time-asymmetry with respect to cosmic time is a typical property of a relevant sub-class of solutions of Einstein's field equations of GR. We take the notion of typicality to characterize a behavior of members of a set that can be shown "to hold with very rare exception".[8] In our case of interest, *most* members of a relevant class of solutions of GR field equations manifest asymmetric behavior, i.e. the asymmetric members of the class build a sub-class of *measure 1* with respect to the Lebesque measure on the solution space with respect to natural co-ordinates provided by the scaling factor, the scalar matter field and their respective derivatives.

The following proof which is based on the work of Castagnino et al. (2003, p. 374f), (Castagnino et al. 2003, p. 990f), see also Wohlfarth 2013) depends on the two following premises:

(i) The concept of cosmic time is not ruled out in the first place via the structure of the considered spacetime (as e.g. via being a non-orientable spacetime).[9]

One could object that (i) would presuppose perfect homogeneity of the universe, while the actual universe is rather inhomogeneous. But the idealization of homogeneity just applies to uniformly co-moving galaxies with respect to which a common time can be defined. Non-uniform motion within the co-moving galaxies thus does not contradict the homogeneity assumption. Given the aim of this paper to show how the asymmetry of causation is based on GR, it seems perfectly adequate to consider the set of spacetimes in which cosmic time can be defined as the relevant reference-set.

[8] cf. (Goldstein 2012, p. 2). Typical behavior is manifested, for example, by the motion of atoms of two metal rods at different temperatures that are brought into thermal contact: the motions evolve so that the temperatures in the rods equalize cf. (Maudlin 2007b, p. 287).

[9] Since the condition of time-orientability guarantees that there is a consistent local time orientation for all points of spacetime, this weaker condition would be sufficient to reduce the class of spacetimes to those that have a unique local time order. But only the stronger condition of the existence of a cosmic time provides a global time function the value of which increases (decreases) along every timelike world line of the universe. Only then we can speak of 'two directions of time' for the whole universe.

The second crucial condition is the following:

(ii) The set of necessary dynamic variables contains further variables apart from the scaling factor.[10]

This condition seems plausible for physical reasons. Even if spacetimes are mathematically possible in which the only dynamic variable is the scaling factor, such spacetimes are unphysical. The set of models (the spacetimes described by a particular solution of Einstein's equation) considered here should be the set of *physically plausible* spacetimes, i.e. the set of spacetimes which contain matter and energy as dynamical entities. Thus, we require that the spacetimes considered here entail a further variable representing their matter and energy content.

Given conditions (i) and (ii), we will show that time-asymmetry with respect to cosmic time is a typical property of spacetimes.

To start with: most of all *open* spacetimes are time-asymmetric. This follows from the fact that we can define the time-asymmetry of open spacetimes according to the asymmetric behaviour of the scaling factor as a function a (t) of the cosmic time t: There exists, for open spacetimes, no hypersurface $t = t_S$ such that for all t: $a(t_S + t) = a(t_S - t)$. Hence, given such open time-orientable spacetime, it is obvious that spacetime geometry looks different in both temporal directions. But, GR also allows *closed* spacetimes. Hence, in order to argue that even in the set of closed time-orientable spacetimes we find cosmic time-asymmetry, we have to analyse closed spacetimes in more detail (cf. (Castagninio et al. 2003, p. 900 f.)).

Our analysis will show that the set of time symmetric spacetimes, even in closed topologies, is a set of measure zero. More precisely, this set of time symmetric and closed solutions of Einstein's equation will turn out to have a lower dimension than the set of all closed solutions.[11]

For simplicity, consider an idealized case where the dynamics of spacetime is described by the scaling factor $a(t)$ and the scalar matter field $\phi(t)$ which depend on cosmic time t. In Hamiltonian mechanics, dynamic equations (the Hamiltonian) depend on dynamic variables and their first derivatives of the cosmic time parameter t. Thus, in this example, we have four variables in the Hamiltonian: $a(t)$, da/dt, $\phi(t)$, $d\phi/dt$. Now, analytic mechanics always allows describing one of these variables as a function of the others. Thus, for simplicity, we have chosen $a(t) = f(da/dt, \phi(t), d\phi/dt)$, where da/dt, $\phi(t)$, $d\phi/dt$ are now independent dynamic variables.

[10]Notice, that Price's use of the Gold universe as a counterexample to any intrinsic time-asymmetry of the universe relies on his considering the scaling factor as the only parameter characterizing the universe.

[11]Spacetimes that have an open but time-symmetric (and not static) topology are open with respect to both past *and* future. We will not consider them because they require a change in the value of the cosmological constant. But, in the context of classical GR, the cosmological constant is *constant* in cosmic time. This may not be the case for full blown quantum or string cosmology, but these yet quite speculative accounts are beyond the scope of this paper. In classical GR a contracting spacetime always includes a Big Crunch (cf. Hawking and Ellis 1973). Thus, a spacetime cannot be open with respect to *two* directions of cosmic time if the spacetime is not static.

If we try to construct a symmetric spacetime, all dynamic variables must together behave in a time-symmetric manner. According to the singularity theorems in classical cosmology (Hawking and Ellis 1973), we know that *a(t)* has just one maximum. Next, we can choose the mathematical origin of cosmic time. For simplicity and without loss of generality, we stipulate the origin so as that *a(0)* is the maximum value of the scaling factor. Thus, as a function of cosmic time, *a(t)* is symmetric in relation to the axis a at the point $t = 0$, i.e. $a|_t = a|_{-t}$. Therefore, it is obvious that *da/dt* is symmetric in relation to the point $(t = 0;\ da/dt = 0)$, i.e. $\mathrm{da/dt}|_t = -\,\mathrm{da/dt}|_{-t}$.

However, for a time-symmetric spacetime, the behaviour of $\phi(t)$ and $d\phi/dt$ *together* with *da/dt* must also be symmetric. Thus, in this example, we have only two possibilities for the behaviour of $\phi(t)$ and $d\phi/dt$ at the cosmic time point $t = 0$, which makes the entire spacetime time-symmetric. These possibilities are given by the triplets $\{da/dt|_{t=0} = 0,\ \phi(t = 0),\ d\phi/dt|_{t=0} = 0\}$, which is a symmetric solution of $\phi(t)$ about the $\phi-$axis at the point $t = 0$, and $\{da/dt|_{t=0} = 0,\ \phi(t = 0) = 0,\ d\phi/dt|_{t=0}\}$ being a point-symmetric solution of $\varphi(t)$ with respect to the point $(t = 0;\ \phi(t = 0) = 0)$: $\varphi|_t = -\,\varphi|_{-t}$.

With respect to the general definition of time-symmetry of spacetimes (Sect. 1), the t = 0-axis represents the spatial hypersurface that splits the whole spacetime into "two halves that are temporal mirror images of each other". It does not matter, for that purpose, whether the respective function is symmetric about the a-axis (hypersurface) at the point t = 0 or whether it is point-symmetric with respect to the point t = 0. In both cases, the respective function develops in the same way for an observer starting at t = 0 and going in the direction of positive values of t as for an observer starting at t = 0 and going in the direction of negative values of t. The spacetime looks physically the same in both of these temporal directions, i.e. it is a symmetric spacetime.

All symmetric solutions can be constructed using the triplets given above. Thus, we can construct a subspace of time-symmetric solutions: *span* $\{(da/dt|_{t=0} = 0,\ \phi(t = 0),\ d\phi/dt|_{t=0} = 0),\ (da/dt|_{t=0} = 0,\ \phi(t = 0) = 0,\ d\phi/dt|_{t=0})\}$. The complete space for solutions of the dynamic equation, however, is given by *span* $\{(da/dt,\ 0,\ 0),\ (0,\ \phi,\ 0),\ (0,\ 0,\ d\phi/dt)\}$

Thus, time-symmetric behaviour of a spacetime occurs only in a subset of solutions having a lower dimension than the entire set of solution *even* if we consider closed spacetimes.

The argument given above has shown that, assuming that cosmic time is definable and that the dynamics of spacetime is described not only by the scaling factor, but by additional variables like the matter field variable, time-asymmetry in terms of cosmic time is a typical property of general relativistic spacetimes which are described in some idealized way. Now, this result would hold true even if more dynamical variables were added, because the same sort of calculation could be applied to more realistically described spacetimes. The entire space of the solutions will turn out to possess higher dimension than the subset of time-symmetric solutions.

This result shows that the fact that the Einstein equations are time reversal invariant does not imply that the *models* of General Relativity are time-symmetric. Rather, as we have shown, *almost all* models, in which cosmic time can be defined, are time-asymmetric. Time-asymmetry is thus a typical property of the relevant set of solutions of the GR field equations.

Now, every time-asymmetric model of GR has a time-reversed model which is also a solution of the Einstein equations. Thus someone could object that even if almost all models are time-asymmetric, the existence of a time-reversed twin-model for each of those time-asymmetric models would destroy any time-directedness for the totality of all models.

But this objection fails.[12] The two models *f(t)* and *f(−t)* can be shown to represent the same physical world. We argue for that in the following three steps:

(a) Since the models are isomorphic, $f(t)$ does not possess intrinsic properties that are not possessed by $f(-t)$.
(b) The models describe spacetime as a whole. This implies that there is no time parameter (or any other physical parameter) outside the range of the geometrical objects $f(t)$ and $f(-t)$. The models are thus not related to a physical environment.
(c) The combination of conditions (a) and (b) shows that two models $f(t)$ and $f(-t)$ do not differ in any intrinsic *or* extrinsic property. Thus, the models represent the same physical world.

It should be noticed that a time-reversed model cannot be distinguished from the original one simply by the fact that observers located within this model would have reversed experiences as compared to the original one (for example, shrinking distances versus extending distances). In order to argue in that way, one would have to presuppose that for the time-reversed as much as for the original model some future (and past) time direction had already been determined. But this presupposition would simply beg the question we are concerned with in that paper.

By now we have only shown that almost all models of GR are globally time-asymmetric (with respect to cosmic time). Since there is no obvious way to connect time-asymmetry of cosmic time with the time behaviour of physical processes, we have not yet shown that those models have the potential to represent asymmetric causal processes.

Hence, in a second step, we argue that the global time-asymmetry of a spacetime has crucial consequences for the proper time parameter of all world lines in that spacetime. More precisely: We shall show that, at least in spacetimes similar to ours, the proper time parameter for individual world lines behaves asymmetrically. This local time-asymmetry, as we shall argue, provides the basis for the representation of causal relations according to GR.

[12] We follow here the argumentation of (Castagnino and Lombardi 2009, p. 18).

3 The Way to Causal Asymmetry

In this Section, we deduce a temporal asymmetry of proper times from the global time-asymmetry[13] and analyse the set of additional assumptions needed to proceed in that way.

Regarding this issue, it is well known (cf. Earman 1974, p. 22) that we can use a non-vanishing, continuous timelike vector field on a time-orientable spacetime to distinguish between the semi-light cones.

In the following, we provide the conditions for the construction of a non-vanishing continuous timelike vector field which, since it describes energy-momentum flow, is a plausible candidate for representing causal connections in spacetime. This mathematical object turns out to have exactly the physical meaning that we desire in order to anchor the concept of causation within fundamental physics. Furthermore, it can be used, according to the method mentioned above, to fix a local time sense. But one should not assume, at this point, that this already gives us the distinction between a local "past" and "future". All that this object provides us with is a method to describe local causal connections that follow a distinct local time sense which can consistently be extended over the whole spacetime – provided that a local time sense had been selected at one point. But what makes the two possible local time senses *physically* distinct?

It is exactly at this point that *global* time-asymmetry comes into play. Given the physical difference between the two cosmic time-directions in almost all spacetimes (cf. Sect. 2) the distinction between the two possible local time senses at a certain point of spacetime also gets a substantial physical meaning: One semi light-cone at the point contains all the timelike vectors pointing in one of two geometrically different cosmic time directions, whereas the other semi light-cone contains all the timelike vectors pointing in the other cosmic time-direction. Thus the global time-asymmetry is transferred to the local realm, with the effect that the local time senses become distinct with respect to the physically distinct global time directions. But notice that the physical distinctiveness still does not give us an answer to the 'which-is-which'-question: We still cannot tell *which of the two local time senses is the 'past' and which is the 'future' time sense*. What we have achieved instead is an answer to the question of how the asymmetry of local causal relations is anchored in global time asymmetry.

The first step in order to construct the desired non-vanishing timelike vector field is now to identify possible physical candidates. As we will see below, for this task it turns out to be useful to consider the energy–momentum tensor:

[13]We follow here the mathematical procedure of (Castagnino et al. 2003, p. 376f; Castagnino and Lombardi 2009, p. 19 f.). But we will not agree with the view of Castagnino et al., according to which positive local energy flow as constructed in this procedure selects a substantial future direction of time and thus defines a local arrow of time.

$$T_{\mu\nu} = (1/8p)\left(R_{\mu\nu}(g) - (1/2)\, g_{\mu\nu}R(g) - Lg_{\mu\nu}\right)^{14} \quad (1)$$

Since the components of $T_{\mu\nu}$, as they occur in (1), do not play the role of a continuous, non-vanishing timelike vector field in general, we have to add two conditions for $T_{\mu\nu}$, namely

I. $T_{\mu\nu}$ is a type one energy–momentum tensor[15]
II. $T_{\mu\nu}$ satisfies the dominant energy condition $T^{00} \geq |T^{\mu\nu}|$ for any orthonormal basis

If condition I is satisfied, then we can write Eq. (1) in the form

$$T_{\mu\nu} = s_0 V_\mu{}^0 V_\nu{}^0 + \sum_{i=1}^{3} s_i V_\mu{}^i V_\nu{}^i \quad (2)$$

$\{V_\mu{}^0, V_\nu{}^i\}$ being an orthonormal tetrad and, as in the standard notation, $V_\mu{}^0$ being a timelike and $V_\nu{}^i$ a spacelike vector with $i \in \{1,2,3\}$

If condition (II) is satisfied as well, then it follows that $s_0 \geq 0$. Therefore, if s_0 is not zero:

A. Conditions (I) and (II) are fulfilled in almost all world models considered in classical cosmology.
B. $V_\mu{}^0(x)$ (where x represents the spacetime coordinates) is a continuous, non-vanishing timelike vector field. Moreover, $T^{0\mu}$ can be interpreted as the physical energy flow, described by a continuous, non-vanishing timelike vector field.[16]

According to B, in all time-asymmetric spacetimes satisfying the conditions I and II, we find a physical vector field on which the time-*asymmetric causal connection* between events can be based. We will say that events C and E are *causally connected* iff there is a time-asymmetric energy flow from C to E.

It might be objected that the foregoing explication of causal connections implies that the causal asymmetry is to be *defined* by time-asymmetry. Indeed, according to our view the asymmetry of the causal connection is derived from global time-asymmetry that is transferred itself to the local realm. But in contrast to some merely conventional stipulation, this derivation provides the causal asymmetry with specific physical significance: the direction into which the causal 'arrow' points (from C to E) differs substantially from the opposite direction by being aligned to some

[14]Here $R_{\mu\nu}$ is the Ricci tensor, R the Ricci curvature, Λ is the cosmological constant and $g_{\mu\nu}$ the metrical tensor.

[15]This means that the tensor can be described in normal orthogonal coordinates. See Hawking and Ellis (1973) and also Eq. (2) for the mathematical meaning of 'type I' or 'normal' in this context.

[16]This interpretation appears to be canonical in the context of general relativity, but there are exceptions that show that this understanding of $T^{0\mu}$ is not valid in general. The exceptions come into play by considering quantum effects. Critical points are, for example, the Casimir effect or squeezed vacuum or Hawking-evaporation. (see e.g. Visser 1996; Barceló and Visser 2002).

distinguishable cosmological time direction in which the geometry "looks" different as compared with the opposite direction. Thus, the *causal directions are physically distinct from each other.*

What we have established by now may be called a *weak* causal arrow: Causal relations between events have a substantial (not conventional) time-direction that is in line with one of the global time-directions which are substantially (not conventionally) different in virtue of their particular geometrical characteristics. But we get no definite answer to the question *which* global direction is selected as the 'past' (or 'future') direction, with 'past' and 'future' having their common meaning as manifested in daily experience ('fixed past principle'). To get an answer to that question would require being able to solve the problem of a *strong* causal arrow – being able to single out a unique global arrow of time which could then be transferred to the local realm with the effect that the selected directions coincide with the experienced past and future. Instead of this, what we actually have achieved is only a *distinction* between two global time directions without any method to tell which is which. We have thus provided only a solution to the problem of the weak causal arrow – to the problem "of finding a substantial asymmetry of time that allows us to distinguish between both temporal directions." (Castagnino and Lombardi 2009, p. 5)

4 Conclusions

We have shown that most of all spacetimes of GR are globally time-asymmetric. This global time-asymmetry can be transferred to the local light-cone structure. By that reason, a continuous timelike vector field which has the physical meaning of energy flow and can thus be considered to represent causal connections between events, defines a local time sense with respect to physically distinct global time directions. We suggest that this is the physical basis of the asymmetry of the causal relation. Thus, we have defeated the Neo-Russellian claim, according to which the asymmetry of causation cannot be derived from physics. Causation, contrary to those claims, has a prominent place in physics, at least with respect to fundamental causal connections.

References

Barceló, C., & Visser, M. (2002). *Twilight of the energy conditions?* Preprint gr-qc/0205066.

Castagnino, M., & Lombardi, O. (2009). The global non-entropic arrow of time: From global geometrical asymmetry to local energy flow. *Synthese, 169*, 1–25.

Castagnino, M., Lara, L., & Lombardi, O. (2003a). The cosmological origin of time asymmetries. *Classical and Quantum Gravity, 20*, 369–391.

Castagnino, M., Lombardi, O., & Lara, L. (2003b). The global arrow of time as a geometrical property of the universe. *Foundations of Physics, 33*(6), 877–912.

Earman, J. (1974). An attempt to add a little direction to "the problem of the direction of time". *Philosophy of Science, 41*, 15–47.

Frisch, M. (2012). No place for causes? Causal skepticism in physics. *European Journal for the Philosophy of Science, 2*(3), 313–336.

Goldstein, S. (2012). *Typicality and notions of probability in physics*. In: Y. Ben-Menahem & M. Hemmo (Eds.), *Probability in physics* (pp. 59–71). Berlin: Springer.

Hawking, S., & Ellis, G. (1973). *The large scale structure of space-time*. Cambridge: Cambridge University Press.

Maudlin, T. (2007a). *The metaphysics within physics*. Oxford: Oxford University Press.

Maudlin, T. (2007b). What could be objective about probabilities? *Studies in History and Philosophy of Modern Physics, 38*, 275–291.

Price, H. (2007). Causal perspectivalism. In: H. Price, & R. Corry (Eds.), *Causality, physics, and the constitution of reality: Russell's republic revisited* (pp. 250–292). Oxford: Oxford University Press.

Price, H., & Corry, R. (Eds.). (2007). *Causality, physics, and the constitution of reality: Russell's republic revisited*. Oxford: Oxford University Press.

Russell, B. (1912/1913). On the notion of cause. *Proceedings of the Aristotelian Society, 13*: 1–26.

Visser, M. (1996). *Lorentzian wormholes: From Einstein to hawking*. New York: Springer.

Wohlfarth, D. (2013). A new view of fundamentality for time asymmetries in modern physics. In *Proceedings of the EPSA 11 Conference in Athens, October 2011: Recent Progress in Philosophy of Science: Perspectives and Foundational Problems*. New York: Springer.

Local Causality and Complete Specification: A Reply to Seevinck and Uffink

Gábor Hofer-Szabó

1 Introduction

Local causality is the idea that causal processes propagate though space continuously and with velocity less than the speed of light. John Stewart Bell formulates this intuition in a 1988 interview as follows:

> [Local causality] is the idea that what you do has consequences only nearby, and that any consequences at a distant place will be weaker and will arrive there only after the time permitted by the velocity of light. Locality [= local causality] is the idea that consequences propagate continuously, that they don't leap over distances. (Mann and Crease 1988)

Bell has returned to this intuitive idea of local causality from time to time and provided a more and more elaborate formulation of it. First he addressed the notion of local causality in his "The theory of local beables" delivered at the Sixth GIFT Seminar in 1975 (Bell 1975); later in a footnote added to his 1986 paper "EPR correlations and EPW distributions" (Bell 1986) intending to clean up the first version; and finally in the most elaborate form in his "La nouvelle cuisine" posthumously published in 1990. In this latter paper local causality obtains the following formulation[1]:

> A theory will be said to be locally causal if the probabilities attached to values of local beables in a space-time region V_A are unaltered by specification of values of local beables in a space-like separated region V_B, when what happens in the backward light cone of V_A

[1] For the sake of conformity with the rest of the paper I slightly changed Bell's notation and figure.

G. Hofer-Szabó (✉)
Research Center for the Humanities, Budapest, Hungary
e-mail: szabo.gabor@btk.mta.hu

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_16

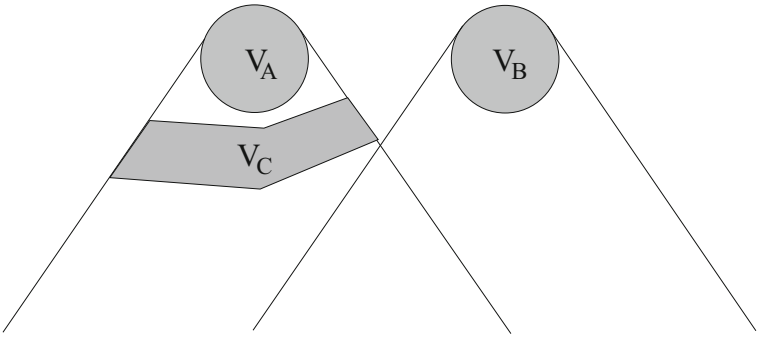


Fig. 1 Full specification of what happens in V_C makes events in V_B irrelevant for predictions about V_A in a locally causal theory

> is already sufficiently specified, for example by a full specification of local beables in a space-time region V_C. (Bell 1990/2004, pp. 239–240)

We reproduce the figure Bell is attaching to his formulation in Fig. 1. (The captation is Bell's original.)

Some brief remarks concerning Bell's terminology are in place here (for a detailed analysis see Norsen 2009, 2011):

(i) The term "beable" in the quote is Bell's own neologism and is contrasted to the term "observable" used in quantum theory. "The *be*ables of the theory are those entities in it which are, at least tentatively, to be taken seriously, as corresponding to something real" (Bell 1990/2004, p. 234).
(ii) Beables are to be local: "*Local* beables are those which are definitely associated with particular space-time regions. The electric and magnetic fields of classical electromagnetism, $\mathbf{E}(t, x)$ and $\mathbf{B}(t, x)$ are again examples." (p. 234).
(iii) Local beables in region V_C are to be "fully specified" in order to block causal influences arriving at V_A from the common past of V_A and V_B.

This latter point is of central importance and is also stressed by Bell[2]:

> It is important that region V_C completely shields off from V_A the overlap of the backward light cones of V_A and V_B. And it is important that events in V_C be specified completely. Otherwise the traces in region V_B of causes of events in V_A could well supplement whatever else was being used for calculating probabilities about V_A. The hypothesis is that any such information about V_B becomes redundant when V_C is specified completely. (Bell 1990/2004, p. 240)

In a recent paper Michael Seevinck and Jos Uffink (2011) have questioned the necessary role of *complete* specification in the definition of local causality and recommended *sufficient* specification instead. They argue that complete specification

[2]But, to be fair, see Bell (1981/2004, p. 106), Bell (1981/2004, p. 152) and the above (Bell 1990/2004, p. 234) for Bell's hesitation on the issue.

is too strong: it contradicts to the so-called no-conspiracy (free variable) condition which requires that the common cause of the correlation be probabilistically independent of the choice of the measurement settings.

I do not see this contradiction and my aim in this paper is to articulate my point. I will proceed as follows. The logical schema of Bell's definition of local causality is the following: if events are localized in the spacetime in such-and-such a way, then these events are to satisfy such-and-such probabilistic independencies. This schema is highly intuitive and easily applicable in the physical praxis, however, in order to account for these inferences from spatiotemporal to probabilistic relations in a mathematically transparent way, one needs to have a *framework* integrating both spatiotemporal and also probabilistic entities. Only after having such a common framework can one define Bell's notion of local causality in a clear-cut way. Thus, in Sect. 2 first this framework, called *local physical theory*, will be introduced and then Bell's notion of local causality will be formulated within this framework. In Sect. 3 the relation of local causality to the Bell inequalities will be explicated. The main section is Sect. 4; here it will be argued that there is no tension between complete specification and no-conspiracy. I conclude in Sect. 5.

2 Bell's Local Causality in a Local Physical Theory

In developing the notion of a local physical theory one is lead by the following intuitions. A local physical theory is to describe "beables," let them be classical or nonclassical; it is to account for the logical combination of these events; these events should be capable of bearing a probabilistic interpretation; the theory is to provide some way to localize these event in the spacetime, and is also to provide some physically well-motivated principles guiding the association of spacetime regions to physical events; the theory is to guarantee that the symmetries of the spacetime are in tune with the symmetries of the events. (For the details see Hofer-Szabó and Vecsernyés 2015a,b.) All these preliminary intuitions are captured in the following definition (Haag 1992):

Definition 1. A $\mathcal{P}_{\mathcal{K}}$*-covariant local physical theory* is a net $\{\mathcal{A}(V), V \in \mathcal{K}\}$ associating algebras of events to spacetime regions which satisfies *isotony*, *microcausality* and *covariance* defined as follows:

1. *Isotony.* Let $\mathcal{M}$ be a globally hyperbolic spacetime and let $\mathcal{K}$ be a covering collection of bounded, globally hyperbolic subspacetime regions of $\mathcal{M}$ such that $(\mathcal{K}, \subseteq)$ is a directed poset under inclusion $\subseteq$. The net of local observables is given by the isotone map $\mathcal{K} \ni V \mapsto \mathcal{A}(V)$ to unital C^*-algebras, that is $V_1 \subseteq V_2$ implies that $\mathcal{A}(V_1)$ is a unital C^*-subalgebra of $\mathcal{A}(V_2)$. The *quasilocal algebra* $\mathcal{A}$ is defined to be the inductive limit C^*-algebra of the net $\{\mathcal{A}(V), V \in \mathcal{K}\}$ of local C^*-algebras.

2. *Microcausality* (also called as *Einstein causality*) is the requirement that $\mathcal{A}(V')' \cap \mathcal{A} \supseteq \mathcal{A}(V), V \in \mathcal{K}$, where primes denote spacelike complement and algebra commutant, respectively.
3. *Spacetime covariance.* Let $\mathcal{P}_\mathcal{K}$ be the subgroup of the group $\mathcal{P}$ of geometric symmetries of $\mathcal{M}$ leaving the collection $\mathcal{K}$ invariant. A group homomorphism $\alpha\colon \mathcal{P}_\mathcal{K} \to \mathrm{Aut}\,\mathcal{A}$ is given such that the automorphisms $\alpha_g, g \in \mathcal{P}_\mathcal{K}$ of $\mathcal{A}$ act covariantly on the observable net: $\alpha_g(\mathcal{A}(V)) = \mathcal{A}(g \cdot V), V \in \mathcal{K}$.

If the quasilocal algebra $\mathcal{A}$ of the local physical theory is commutative, we speak about a *local classical theory*, if it is noncommutative, we speak about a *local quantum theory*. For local classical theories microcausality fulfills trivially.

A *state* ϕ in a local physical theory is defined as a normalized positive linear functional on the quasilocal observable algebra $\mathcal{A}$. The corresponding GNS representation $\pi_\phi\colon \mathcal{A} \to \mathcal{B}(\mathcal{H}_\phi)$ converts the net of C^*-algebras into a net of C^*-subalgebras of $\mathcal{B}(\mathcal{H}_\phi)$. Closing these subalgebras in the weak topology one arrives at a net of local von Neumann observable algebras: $\mathcal{N}(V) := \pi_\phi(\mathcal{A}(V))'', V \in \mathcal{K}$. Von Neumann algebras are generated by their projections, which are called *quantum events* since they can be interpreted as 0-1–valued observables. The net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ of local von Neumann algebras given above also obeys isotony, microcausality, and $\mathcal{P}_\mathcal{K}$-covariance, hence we can also refer to a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ of local von Neumann algebras as a local physical theory.

Now, a local physical theory is locally causal in Bell's sense if any correlation between spatially separated events is screened off by "local beables" which are localized in a "shielding-off" region and which "completely specify" that region. How to translate Bell's terms of "local beable" and "complete specification" into the language of a local physical theory?

In a classical field theory beables are characterized by sets of field configurations. Taking the equivalence classes of those field configurations which have the same field values on a given spacetime region one can generate local (cylindrical) σ-algebras. Translating σ-algebras into the language of abelian von Neumann algebras one can represent Bell's notion of "local beables" in the framework of local physical theories. In a more general way, one can also use the term "local beables" both for abelian and non-abelian local von Neumann algebras, hence treating local classical and quantum theories on an equal footing. Translating "local beables" simply as "elements of a local algebra" naturally brings with it the translation of the term "a complete specification of beables" as "an atomic event of a local algebra" (Henson 2013). To be sure, here it is assumed that the local algebras of the net are atomic, which is typically not the case, for example, in Poincaré covariant algebraic quantum field theory. (For the relation between σ-algebras and von Neumann algebras and for a more general definition of local causality see Hofer-Szabó and Vecsernyés 2015a,b.) With these notions in hand now one can formulate Bell's notion of local causality in a local physical theory as follows:

Definition 2. A local physical theory represented by a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ of von Neumann algebras is called *locally causal* (in Bell's sense), if for any pair

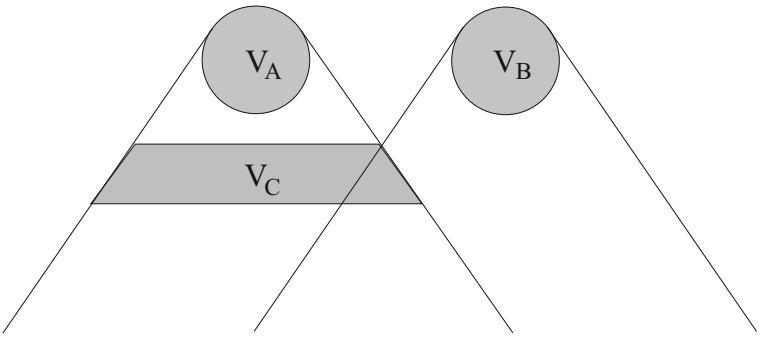


Fig. 2 A region V_C satisfying Requirements (i)–(iii)

$A \in \mathcal{N}(V_A)$ and $B \in \mathcal{N}(V_B)$ of projections supported in spacelike separated regions $V_A, V_B \in \mathcal{K}$ and for every locally normal and faithful state ϕ establishing a correlation $\phi(AB) \neq \phi(A)\phi(B)$ between A and B, and for any spacetime region V_C such that

(i) $V_C \subset J_-(V_A)$,
(ii) $V_A \subset V_C''$,
(iii) $J_-(V_A) \cap J_-(V_B) \cap \left(J_+(V_C) \setminus V_C\right) = \emptyset$,

(see Fig. 2) and for any atomic event C_k of $\mathcal{A}(V_C)$ ($k \in K$), the following holds:

$$\frac{\phi(C_k A B C_k)}{\phi(C_k)} = \frac{\phi(C_k A C_k)}{\phi(C_k)} \frac{\phi(C_k B C_k)}{\phi(C_k)} \tag{1}$$

Remarks. 1. A *locally normal* state is a normal state on the local von Neumann algebras. A *locally faithful* state ϕ means that any projection $A \in \mathcal{P}(\mathcal{N}(V))$ in the local von Neumann algebra $\mathcal{N}(V)$ has nonzero expectation value. In case of local classical theories a locally faithful state ϕ determines uniquely a locally nonzero probability measure p by $p(A) := \phi(A), A \in \mathcal{P}(\mathcal{N}(V))$. By means of this (1) can be written in the following 'symmetric' form:

$$p(AB|C_k) = p(A|C_k)p(B|C_k) \tag{2}$$

which further is equivalent to the 'asymmetric' screening-off condition:

$$p(A|BC_k) = p(A|C_k) \tag{3}$$

sometimes used in the literature (for example in Bell 1975/2004, p. 54).

2. The role of Requirement (iii) in the definition is to ensure that "V_C shields off from V_A the overlap of the backward light cones of V_A and V_B". A spacetime region *above* V_C in the common past of the correlating events (see Fig. 3) namely may contain stochastic events which could establish a correlation between A and

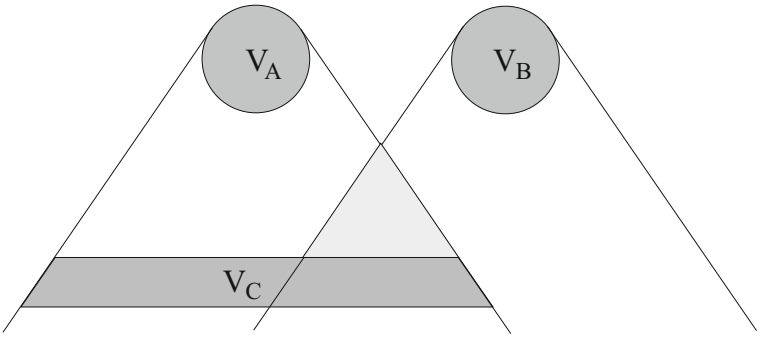


Fig. 3 A region V_C for which Requirement (iii) does not hold

B in a classical stochastic theory (Norsen 2011; Seevinck and Uffink 2011). Requirement (iii) is somewhat weaker than Bell's original localization (see Fig. 1) which can be formulated as:

(iv) $J_-(V_A) \cap J_-(V_B) \cap V_C = \emptyset$

The difference is that Requirement (iii) does, but Requirement (iv) does not allow for region V_C to penetrate into the 'top part' of the common past. However, both requirements coincide, if V_C 'shrinks down' to a Cauchy surface. In local classical theories it suffices to use Requirement (iii).

Finally, note that the question whether a given local classical or quantum theory is locally causal is a highly nontrivial question depending on such factors as the atomicity of the local algebras, the fulfilment of the so-called local primitive causality,[3] or whether there exists a causal dynamics in the theory, etc. (For the details see again Hofer-Szabó and Vecsernyés 2015a,b.)

Next I turn to the relation of Bell's local causality to the Bell inequalities.

3 Local Causality and the Bell Inequalities

From this section on we restrict ourselves to local *classical* theories since beables are standardly taken to be classical entities. Consider a local classical theory represented by a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ of local abelian von Neumann algebras. Suppose that Bell's local causality holds in this theory. Let V_A and V_B be two spatially separated regions in $\mathcal{M}$, and V_C a third region (see Fig. 4) such that

$$V_C \subset J_-(V_A \cup V_B) \tag{4}$$

[3]For any globally hyperbolic bounded subspacetime regions $V \in \mathcal{K}$, $\mathcal{A}(V'') = \mathcal{A}(V)$.

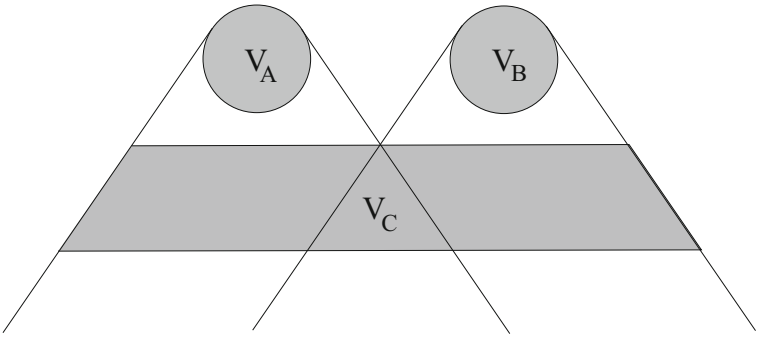


Fig. 4 Localization of regions V_A, V_B and V_C

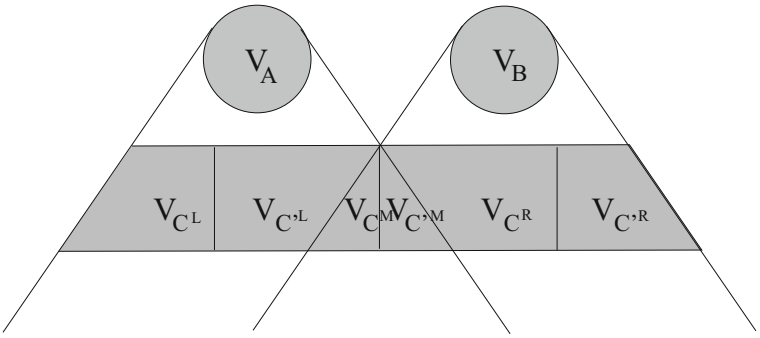


Fig. 5 Dividing up region V_C

$$(V_A \cup V_B) \subset V_C'' \tag{5}$$

$$J_-(V_A) \cap J_-(V_B) \cap \big(J_+(V_C) \setminus V_C\big) = \emptyset \tag{6}$$

Divide V_C into six regions V_C^L, $V_{C'}^L$, V_C^M, $V_{C'}^M$, V_C^R and $V_{C'}^R$, for example as depicted in Fig. 5. Here the superscripts L, M and R stand for 'left', 'middle' and 'right', representing those parts of V_C which fall into region $J_-(V_A) \setminus J_-(V_B)$, $J_-(V_A) \cap J_-(V_B)$ and $J_-(V_B) \setminus J_-(V_A)$, respectively. Now, let the various events be localized in these regions as follows. Let A_i and B_j be *measurement outcomes* and a_i, b_j *measurement choices* localized in the appropriate regions: $A_i, a_i \in \mathcal{A}(V_A)$, $B_j, b_j \in \mathcal{A}(V_B)$. (See Fig. 6.) Let, furthermore, C_k^L, $C_l'^L$, C_m^M, $C_n'^M$, C_p^R, $C_q'^R$ be *atomic* events (minimal projections) in $\mathcal{A}(V_C^L)$, $\mathcal{A}(V_{C'}^L)$, $\mathcal{A}(V_C^M)$, $\mathcal{A}(V_{C'}^M)$, $\mathcal{A}(V_C^R)$ and $\mathcal{A}(V_{C'}^R)$, respectively, where the indices $i, j, k \ldots$ are taken from appropriate index sets. Now, the difference between the primed and the unprimed events in V_C is that the *primed* events probabilistically *depend* on the measurement choices a_i and b_j, whereas the *unprimed* events are probabilistically *completely independent* of them:

$$p(a_i b_j C_l^L C_m^M C_p^R) \;=\; p(a_i)p(b_j)p(C_l^L)p(C_m^M)p(C_p^R) \tag{7}$$

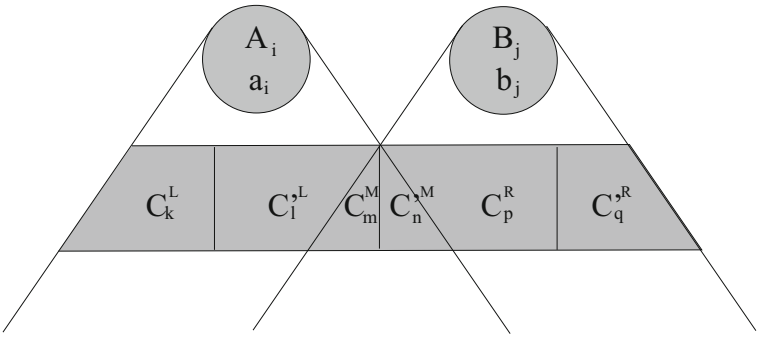


Fig. 6 Localization of the various events

$$p(a_i b_j C_l^L C_m^M) = p(a_i)p(b_j)p(C_l^L)p(C_m^M) \tag{8}$$

$$\ldots \tag{9}$$

$$p(a_i b_j C_p^R) = p(a_i)p(b_j)p(C_p^R) \tag{10}$$

Let us call these conditions *no-conspiracy conditions*.

To sum up, here we assume that *any* of the left, middle and right region of V_C, respectively can be decomposed into two subregions such that each of these subregions contains exclusively *either* events 'influencing' the measurement choices *or* events being independent of them. Obviously, only this latter class of events can be regarded as the *common cause* of the correlation between the measurement outcomes; the former events are playing a role in fixing the measurement settings. As we will see later, this assumption of the decomposability of V_C into *six* regions is too tolerant if our aim is to derive the Bell inequalities. It will turn out that there are only *five* regions, the middle region can contain only *unprimed* events.

Now, *local causality* of local physical theory represented by a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ implies (among others) the following conditional independencies:

$$p(A_i a_i | B_j b_j C_k^L C_l'^L C_m^M C_n'^M C_p^R C_q'^R) = p(A_i a_i | C_k^L C_l'^L C_m^M C_n'^M) \tag{11}$$

$$p(B_j b_j | C_k^L C_l'^L C_m^M C_n'^M C_p^R C_q'^R) = p(B_j b_j | C_m^M C_n'^M C_p^R C_q'^R) \tag{12}$$

$$p(a_i | b_j C_k^L C_l'^L C_m^M C_n'^M C_p^R C_q'^R) = p(a_i | C_k^L C_l'^L C_m^M C_n'^M) \tag{13}$$

$$p(b_j | C_k^L C_l'^L C_m^M C_n'^M C_p^R C_q'^R) = p(b_j | C_m^M C_n'^M C_p^R C_q'^R) \tag{14}$$

which together with the *complete independence* of the events C_k^L, $C_l'^L$, C_m^M, $C_n'^M$, C_p^R and $C_q'^R$:

$$p(C_k^L C_l'^L C_m^M C_n'^M C_p^R C_q'^R) = p(C_k^L)p(C_l'^L)p(C_m^M)p(C_n'^M)p(C_p^R)p(C_q'^R) \tag{15}$$

$$p(C_k^L C_l'^L C_m^M C_n'^M C_p^R) = p(C_k^L)p(C_l'^L)p(C_m^M)p(C_n'^M)p(C_p^R) \quad (16)$$

$$\ldots \quad (17)$$

$$p(C_p^R C_q'^R) = p(C_p^R)p(C_q'^R) \quad (18)$$

yield the following *screening-off* or *factorization conditions*:

$$p(A_iB_j|a_ib_jC_k^L C_l'^L C_m^M C_n'^M C_p^R C_q'^R) = p(A_i|a_iC_k^L C_l'^L C_m^M C_n'^M)p(B_j|b_jC_m^M C_n'^M C_p^R C_q'^R) \quad (19)$$

$$p(A_iB_j|a_ib_jC_k^L C_m^M C_n'^M C_p^R) = p(A_i|a_iC_k^L C_m^M C_n'^M)p(B_j|b_jC_m^M C_n'^M C_p^R) \quad (20)$$

$$p(A_iB_j|a_ib_jC_l'^L C_m^M C_n'^M C_q'^R) = p(A_i|a_iC_l'^L C_m^M C_n'^M)p(B_j|b_jC_m^M C_n'^M C_q'^R) \quad (21)$$

$$p(A_iB_j|a_ib_jC_m^M C_n'^M) = p(A_i|a_iC_m^M C_n'^M)p(B_j|b_jC_m^M C_n'^M) \quad (22)$$

(For the proof see Appendix 1.) These equations show that not only the *atomic* events $C_k^L C_l'^L C_m^M C_n'^M C_p^R C_q'^R$ localized in the entire V_C screen off the conditional correlation

$$p(A_iB_j|a_ib_j) \neq p(A_i|a_i)p(B_j|b_j) \quad (23)$$

but one can freely sum up for any of the *primed* and *unprimed* events both in the *left* and the *right* region without vitiating the screening-off. In other words, the non-atomic (coarse-grained) events $C_k^L C_m^M C_n'^M C_p^R$, $C_l'^L C_m^M C_n'^M C_q'^R$ and $C_m^M C_n'^M$, respectively localized in appropriate subregions of V_C will all be screener-offs for the correlation (23).[4] That one can freely sum up for both the primed and the unprimed events is a consequence of the fact that in the derivation of (19), (20), (21), and (22) no-conspiracy (7), (8), (9), and (10) does *not* play a role.

However, for events localized in the *middle* region one *cannot* sum up! As a consequence, one cannot get rid of the *primed* terms $C_n'^M$ in Eqs. (19), (20), (21), and (22). So for example it will *not* be generally true that

$$p(A_iB_j|a_ib_jC_m^M) = p(A_i|a_iC_m^M)p(B_j|b_jC_m^M) \quad (24)$$

(See Appendix 2.) However, if we cannot get rid of the *primed* terms $C_n'^M$, we will not be able to derive the Bell inequalities since in the derivation we need to use no-conspiracy (7), (8), (9), and (10) which holds only for the *unprimed* terms. (See Appendix 3.)

This shows that our decomposition of region V_C into *six* regions was too liberal. We have to make one step back and restrict our previous schema to the one depicted in Fig. 7. *Outside* the common past of the correlating events one can have both primed and unprimed events that is events influencing the measurement choices

[4]Note again that the term 'common cause' is used only for those screener-offs which are composed of *unprimed* events.

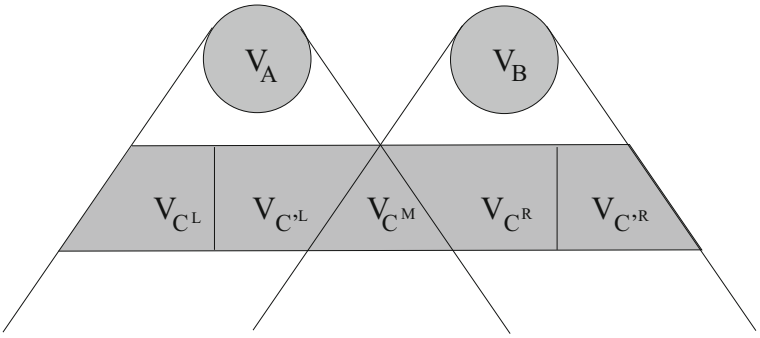


Fig. 7 The most general scenario from which the Bell inequalities can be derived

and events being independent of them. However, *within* the common past there can be only events which are probabilistically independent of the measurement choices. Within this schema the Bell inequalities can be derived.

To sum up, given a locally causal local classical theory represented by a net $\{\mathcal{N}(V), V \in \mathcal{K}\}$ with regions localized as in Fig. 7 and elements in the appropriate regions, complete independence (15), (16), (17), and (18) and no-conspiracy (7), (8), (9), and (10) together imply the Bell inequalities.

4 Complete *Versus* Sufficient Specification

Now I turn to the question of 'complete *versus* sufficient specification' raised by Norsen (2009) and unfolded by Seevinck and Uffink (2011). In his illuminating paper, comparing the notion of 'completeness' used in Bell's *vs*. Jarrett's writings, Norsen (2009) raised the following concern[5]: Since "the past light cones of [the measurement choices] a and b overlap with the region containing C – and C by definition is supposed to contain a *complete* specification of beables in this region ... one wonders how a and b could possibly *not* be causally influenced by C (in a locally causal theory)" (Norsen 2009, p. 283.) Seevinck and Uffink take Norsen's point and argue that complete specification is too strong "when formalising the notion of local causality. It is only needed that the specification is *sufficiently* specified, in the relevant sense" (p. 5); and then they go on to develop this relevant sense in terms of Fisher's statistical concept of sufficiency.

The argument of Seevinck and Uffink against complete specification is put in the form of a dilemma:

[5] Again for the sake of consistency I changed the notation of both Norsen (2009) and Seevinck and Uffink (2011).

> C cannot be expected to be a complete specification of region V_C because one must allow for the possibility of traces in region V_C of the causal past of both the settings [measurement choices], and given the independence of C and the settings, these traces cannot be included in C.
>
> An alternative understanding of this point is that one is here faced with a dilemma. That is, the following two assumptions cannot both hold: (i) the free variables [no-conspiracy] assumption, and (ii) the assumption that C is completely specified, i.e., contains the description of all and every beable in region V_C. (Seevinck and Uffink 2011, p. 5)

In brief, the complete specification of region V_C contradicts to the no-conspiracy condition since if C completely specifies region V_C, then it also specifies the measurement choices a and b, and hence C and a, b cannot be probabilistically independent.

I see, however, no contradiction between complete specification and no-conspiracy. I have a weaker and a stronger claim supporting my point. I start with the weaker one. The upshot of this weaker claim is that the events which satisfy complete specification *need not be the same* as the events which satisfy no-conspiracy.

Complete specification of a spacetime region, as said before, is simply an atomic event in that region. If our "candidate theory" represented by a net of local algebras is given, then to every bounded region V_C there is an algebra $\mathcal{A}(V_C)$ associated; and if the algebra is atomic, the complete specifications that is the atomic events of the region are also given. Consider region V_C in Fig. 7. The event $C_k^L C_l'^L C_m^M C_p^R C_q'^R$ is a complete specification in V_C, but the unprimed event $C_k C_m C_p$ and the primed event $C_l' C_q'^R$ separately are not. These latter two play different theoretical roles: No-conspiracy holds for $C_k C_m C_p$, therefore it is interpreted as a (possible) common cause of the conditional correlation (23). For $C_l' C_q'^R$ no-conspiracy does not hold (and a fortiori neither does for the complete specification $C_k^L C_l'^L C_m^M C_p^R C_q'^R$). Thus $C_l' C_q'^R$ has another interpretation: it allows "for the possibility of traces in region V_C of the causal past of both the settings." This 'division of labor' between the unprimed $C_k C_m C_p$ and the primed $C_l' C_q'^R$, however, is no worry: together they provide a complete specification of region V_C and enable the derivation of the Bell inequalities as long as the middle region, $V_C \cap V_A \cap V_B$ contains no primed term violating no-conspiracy. In short, in order to derive the Bell inequalities from local causality, those events which completely specify region V_C *need not be the same events* as those satisfying no-conspiracy.

But here is my stronger claim: *they can*. Namely, there is no contradiction between complete specification and no-conspiracy even if we require them to hold *for the same events*. To see this, simply consider the case when the subregions $V_{C'}^L$ and $V_{C'}^R$ are empty, that is when V_C contains exclusively unprimed elements (see Fig. 8). In this case the event $C_k^L C_m^M C_p^R$ will *both* completely specify region V_C *and* satisfy no-conspiracy. Consequently, the Bell inequalities will follow. More importantly, this independence between the common causes and the measurement choices does not trivialize the theory, for example by dissolving the conditional correlation (23) between the measurement outcomes.

The next proposition illustrates this latter point.

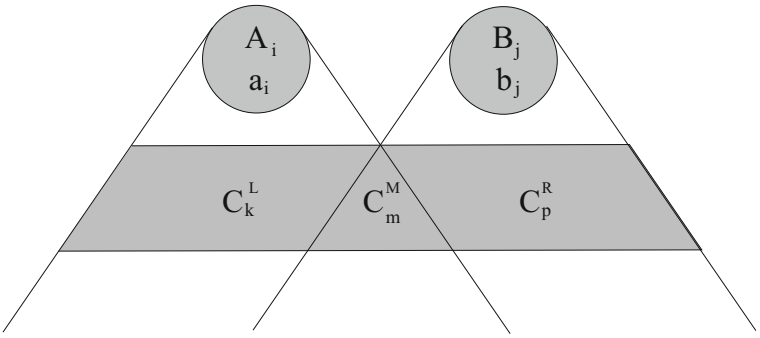


Fig. 8 No contradiction between complete specification and no-conspiracy

Proposition 1. *There exists a locally causal local classical theory with events* $A_i, a_i \in \mathcal{A}(V_A)$, $B_j, b_j \in \mathcal{A}(V_B)$ *in spatially separated regions* V_A *and* V_B *conditionally correlating in the sense of (23), and atomic events* $C_k^L \in \mathcal{A}(V_C^L)$, $C_m^M \in \mathcal{A}(V_C^M)$ *and* $C_p^R \in \mathcal{A}(V_C^R)$, *where* $V_C = V_C^L \cup V_C^M \cup V_C^R$ *satisfies requirements (4), (5), and (6), such that no-conspiracy (7), (8), (9), and (10), moreover complete independence (15), (16), (17), and (18) hold.*

Proof. Let A_i, a_i, B_j, b_j, C_k^L, C_m^M and C_p^R be events localized as in Fig. 8. Suppose that for the atomic events C_k^L, C_m^M and C_p^R completely specifying region V_C both complete independence

$$\begin{aligned} p(C_k^L C_m^M C_p^R) &= p(C_k^L C_m^M)p(C_p^R) = p(C_k^L)p(C_m^M C_p^R) = p(C_m^M)p(C_k^L C_p^R) \\ &= p(C_k^L)p(C_m^M)p(C_p^R) \end{aligned} \tag{25}$$

and also no-conspiracy

$$p(a_i b_j C_k^L C_m^M C_p^R) = p(a_i b_j)p(C_k^L C_m^M C_p^R) = \cdots = p(a_i)p(b_j)p(C_k^L)p(C_m^M)p(C_p^R) \tag{26}$$

hold for any combination of the indices. Let the net containing the events be locally causal; for example let

$$p(A_i B_j | a_i b_j C_k^L C_m^M C_p^R) = p(A_i | a_i C_k^L C_m^M)p(B_j | b_j C_m^M C_p^R) = (p_i^L \delta_{1k}\delta_{1m})(p_j^R \delta_{1m}\delta_{1p}) \tag{27}$$

where $\sum_i p_i^L = \sum_j p_j^R = 1$. Now, the conditional probabilities are given as follows:

$$p(A_i|a_i) = \sum_{k,m} p(A_i | a_i C_k^L C_m^M)p(C_k^L C_m^M) = p_i^L p(C_1^L)p(C_1^M) \tag{28}$$

$$p(B_j|b_j) = \sum_{m,p} p(B_j | b_j C_m^M C_p^R)p(C_m^M C_p^R) = p_j^R p(C_1^M)p(C_1^R) \tag{29}$$

$$
\begin{aligned}
p(A_iB_j|a_ib_j) &= \sum_{k,m,p} p(A_iB_j|a_ib_jC_k^LC_m^MC_p^R)p(C_k^LC_m^MC_p^R) \\
&= \sum_{k,m,p} p(A_i|a_iC_k^LC_m^M)p(B_j|b_jC_m^MC_p^R)p(C_k^L)p(C_m^M)p(C_p^R) \\
&= p_i^Lp_j^Rp(C_1^L)p(C_1^M)p(C_1^R) \qquad (30)
\end{aligned}
$$

Thus, there is a conditional correlation (23) between A_i and B_j whenever $p(C_1^M) \neq 0$ or 1. ■

Consequently, there is no contradiction between complete specification and no-conspiracy even if both are applied to the same events, namely the atomic events of the entire V_C. The measurement choices can be free of the common causes even if the causal past of the region containing them is completely specified. This independence does not abolish the conditional correlation between the measurement outcomes: atomic events can be probabilistically irrelevant to the measurement *choices* and at the same time relevant to the measurement *outcomes*. Moreover, the independence of the measurement choices of the atomic events does not mean that the former are not 'determined' (probabilistically) by the latter. They are: the conditional probabilities $p(a_ib_j|C_k^LC_m^MC_p^R)$ are set in a local physical theory, even if they are equal to $p(a_ib_j)$.

Thus, based on these two claims, I think, there is no need to replace 'complete specification' in Bell's definition of local causality by 'sufficient specification'.

5 Conclusions

The main claims of this paper were the following:

(i) The definition of Bell's notion of local causality presupposes a clear-cut framework in which probabilistic and spatiotemporal entities can be related. This goal can be met by introducing the notion of a local physical theory represented by an isotone net of algebras.

(ii) In a local classical theory the measurement outcomes, measurement choices and common cause can be localized in the spacetime such that one can derive the Bell inequalities from local causality, no-conspiracy and independence.

(iii) Contrary to the claim of Seevinck and Uffink, there is no need to weaken the requirement of complete specification in the definition of local causality on the ground that it contradicts to no-conspiracy.

Appendix 1

First we prove Eq. (22) from local causality (11), (12), (13), and (14) and the complete independence condition (15), (16), (17), and (18):

$$
\begin{aligned}
p(A_iB_j|a_ib_jC_m^MC_n'^M) &= \frac{p(A_iB_ja_ib_jC_m^MC_n'^M)}{p(a_ib_jC_m^MC_n'^M)} \\
&= \frac{\sum_{klpq} p(A_iB_ja_ib_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klpq} p(a_ib_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
&= \frac{\sum_{klpq} p(A_iB_ja_ib_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klpq} p(a_ib_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
&= \frac{\sum_{klpq} p(A_ia_i|B_jb_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(B_jb_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klpq} p(a_i|b_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(b_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
&\overset{(11)-(14)}{=} \frac{\sum_{klpq} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klpq} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
&\overset{(15)-(18)}{=} \frac{\sum_{klpq} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^M)p(C_p^RC_q'^R)}{\sum_{klpq} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^M)p(C_p^RC_q'^R)} \\
&= \left(\frac{\sum_{kl} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}{\sum_{kl} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}\right)\left(\frac{\sum_{pq} p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_p^RC_q'^R)}{\sum_{pq} p(b_j|C_m^MC_n'^MC_p^RC_q'^RC_p^R)p(C_p^RC_q'^R)}\right) \\
&= \left(\frac{\sum_{kl} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}{\sum_{kl} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}\right)\left(\frac{\sum_{pq} p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_p^RC_q'^R)}{\sum_{pq} p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_p^RC_q'^R)}\right) \\
&\quad \times\left(\frac{p(C_m^MC_n'^M)}{p(C_m^MC_n'^M)}\right) \\
&\overset{(15)-(18)}{=} \left(\frac{\sum_{kl} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}{\sum_{kl} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}\right) \\
&\quad \times\left(\frac{\sum_{pq} p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_m^MC_n'^MC_p^RC_q'^R)}{\sum_{pq} p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_m^MC_n'^MC_p^RC_q'^R)}\right) \\
&= \left(\frac{\sum_{kl} p(A_ia_iC_k^LC_l'^LC_m^MC_n'^M)}{\sum_{kl} p(a_iC_k^LC_l'^LC_m^MC_n'^M)}\right)\left(\frac{\sum_{pq} p(B_jb_jC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{pq} p(b_jC_m^MC_n'^MC_p^RC_q'^R)}\right) \\
&= \left(\frac{p(A_ia_iC_m^MC_n'^M)}{p(a_iC_m^MC_n'^M)}\right)\left(\frac{p(B_jb_jC_m^MC_n'^M)}{p(b_jC_m^MC_n'^M)}\right) = p(A_i|a_iC_m^MC_n'^M)p(B_j|b_jC_m^MC_n'^M) \qquad (31)
\end{aligned}
$$

where the numbers over the equation signs refer to the equation used at that step.

The proof of (21), (20) and (19), respectively can be obtained from the above proof by simply omitting certain summations. For (21) just omit summation for l and r; for (20) omit summation for k and q; and for (19) omit all four.

Appendix 2

Here we prove that (24) does not generally hold. The proof follows that in Appendix 1, except that here there is an extra summation also for n, which causes the trouble in the row below starting with a $\neq$ sign:

$$
\begin{aligned}
& p(A_iB_j|a_ib_jC_m^M) = \frac{p(A_iB_ja_ib_jC_m^M)}{p(a_ib_jC_m^M)} \\
= \quad & \frac{\sum_{klnpq} p(A_iB_ja_ib_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klnpq} p(a_ib_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
= \quad & \frac{\sum_{klnpq} p(A_iB_ja_ib_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klnpq} p(a_ib_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
= \quad & \frac{\sum_{klnpq} p(A_ia_i|B_jb_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(B_jb_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klnpq} p(a_i|b_jC_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(b_j|C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
\overset{(11)-(14)}{=} \quad & \frac{\sum_{klnpq} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{klnpq} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^MC_p^RC_q'^R)} \\
\overset{(15)-(18)}{=} \quad & \frac{\sum_{klnpq} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^M)p(C_p^RC_q'^R)}{\sum_{klnpq} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^M)p(C_p^RC_q'^R)} \\
= \quad & \frac{\sum_n \left(\sum_{kl} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)\sum_{pq} p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_p^RC_q'^R)\right)}{\sum_n \left(\sum_{kl} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)\sum_{pq} p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_p^RC_q'^R)\right)} \\
= \quad & \frac{\sum_n \left(\sum_{kl} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)\sum_{pq} p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_p^RC_q'^R)\right)}{\sum_n \left(\sum_{kl} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)\sum_{pq} p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_p^RC_q'^R)\right)} \\
& \qquad \times\left(\frac{p(C_m^MC_n'^M)}{p(C_m^MC_n'^M)}\right) \\
\neq \quad & \left(\frac{\sum_{kln} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}{\sum_{kln} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}\right) \\
& \qquad \times\left(\frac{\sum_{npq} p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_m^MC_n'^M)p(C_p^RC_q'^R)}{\sum_{npq} p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_m^MC_n'^M)p(C_p^RC_q'^R)}\right) \\
\overset{(15)-(18)}{=} \quad & \left(\frac{\sum_{kln} p(A_ia_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}{\sum_{kln} p(a_i|C_k^LC_l'^LC_m^MC_n'^M)p(C_k^LC_l'^LC_m^MC_n'^M)}\right) \\
& \qquad \times\left(\frac{\sum_{npq} p(B_jb_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^M)}{\sum_{npq} p(b_j|C_m^MC_n'^MC_p^RC_q'^R)p(C_k^LC_l'^LC_m^MC_n'^M)}\right) \\
= \quad & \left(\frac{\sum_{kln} p(A_ia_iC_k^LC_l'^LC_m^MC_n'^M)}{\sum_{kln} p(a_iC_k^LC_l'^LC_m^MC_n'^M)}\right)\left(\frac{\sum_{npq} p(B_jb_jC_m^MC_n'^MC_p^RC_q'^R)}{\sum_{npq} p(b_jC_m^MC_n'^MC_p^RC_q'^R)}\right) \\
= \quad & \left(\frac{p(A_ia_iC_m^M)}{p(a_iC_m^M)}\right)\left(\frac{p(B_jb_jC_m^M)}{p(b_jC_m^M)}\right) = p(A_i|a_iC_m^M)p(B_j|b_jC_m^M) \qquad (32)
\end{aligned}
$$

where again the numbers over the equation signs refer to the equation used at that step.

Appendix 3

Here we prove why in the derivation of the Clauser-Horne inequality

$$-1 \leqslant p(A_iB_j|a_ib_j) + p(A_iB_{j'}|a_ib_{j'}) + p(A_{i'}B_j|a_{i'}b_j) - p(A_{i'}B_{j'}|a_{i'}b_{j'}) - p(A_i|a_ib_j) - p(B_j|a_ib_j) \leqslant 0 \tag{33}$$

one should use (24) instead of (22). The standard derivation goes as follows:

It is a simple arithmetic fact that for any $\alpha, \alpha', \beta, \beta' \in [0, 1]$:

$$-1 \leqslant \alpha\beta + \alpha\beta' + \alpha'\beta - \alpha'\beta' - \alpha - \beta \leqslant 0 \tag{34}$$

Now let $\alpha, \alpha', \beta, \beta'$ first be the conditional probabilities taken from (22):

$$\alpha \equiv p(A_i|a_iC_m^MC_n'^M) \tag{35}$$

$$\alpha' \equiv p(A_{i'}|a_{i'}C_m^MC_n'^M) \tag{36}$$

$$\beta \equiv p(B_j|b_jC_m^MC_n'^M) \tag{37}$$

$$\beta' \equiv p(B_{j'}|b_{j'}C_m^MC_n'^M) \tag{38}$$

Plugging (35), (36), (37), and (38) into (34) one obtains

$$\begin{aligned} -1 \leqslant\ & p(A_i|a_iC_m^MC_n'^M)p(B_j|b_jC_m^MC_n'^M) + p(A_i|a_iC_m^MC_n'^M)p(B_{j'}|b_{j'}C_m^MC_n'^M) \\ & + p(A_{i'}|a_{i'}C_m^MC_n'^M)p(B_j|b_jC_m^MC_n'^M) - p(A_{i'}|a_{i'}C_m^MC_n'^M)p(B_{j'}|b_{j'}C_m^MC_n'^M) \\ & - p(A_i|a_iC_m^MC_n'^M) - p(B_j|b_jC_m^MC_n'^M) \leqslant 0 \end{aligned} \tag{39}$$

which using (22) transforms into

$$\begin{aligned} -1 \leqslant\ & p(A_iB_j|a_ib_jC_m^MC_n'^M) + p(A_iB_{j'}|a_ib_{j'}C_m^MC_n'^M) \\ & + p(A_{i'}B_j|a_{i'}b_jC_m^MC_n'^M) - p(A_{i'}B_{j'}|a_{i'}b_{j'}C_m^MC_n'^M) \\ & - p(A_i|a_iC_m^MC_n'^M) - p(B_j|b_jC_m^MC_n'^M) \leqslant 0 \end{aligned} \tag{40}$$

Finally, multiplying the above inequality by $p(C_m^MC_n'^M)$ and summing up for the indices m, n one obtains

$$-1 \leqslant \sum_{mn} \Big[p(A_i B_j | a_i b_j C_m^M C_n'^M) + p(A_i B_{j'} | a_i b_{j'} C_m^M C_n'^M) + p(A_{i'} B_j | a_{i'} b_j C_m^M C_n'^M) - p(A_{i'} B_{j'} | a_{i'} b_{j'} C_m^M C_n'^M) - p(A_i | a_i C_m^M C_n'^M) - p(B_j | b_j C_m^M C_n'^M) \Big] p(C_m^M C_n'^M) \leqslant 0 \tag{41}$$

which is equivalent to (33) only if

$$p(a_i b_j C_m^M C_n'^M) = p(a_i b_j) p(C_m^M C_n'^M) \tag{42}$$

were the case, which is not, since $C_n'^M$ is not independent of a_i and b_j.

Now, starting the whole reasoning again with conditional probabilities taken from (24):

$$\alpha \equiv p(A_i | a_i C_m^M) \tag{43}$$

$$\alpha' \equiv p(A_{i'} | a_{i'} C_m^M) \tag{44}$$

$$\beta \equiv p(B_j | b_j C_m^M) \tag{45}$$

$$\beta' \equiv p(B_{j'} | b_{j'} C_m^M) \tag{46}$$

the derivation goes through since instead of (42) one is to use

$$p(a_i b_j C_m^M) = p(a_i b_j) p(C_m^M) \tag{47}$$

which is one of the no-conspiracy conditions (7), (8), (9), and (10). Thus one can use (24) in the derivation of the Clauser-Horne inequality but not (22).

Acknowledgements This work has been supported by the Hungarian Scientific Research Fund OTKA K-100715.

References

Bell, J. S. (1975). Beables for quantum field theory. In *TH-2053-CERN, Presented at the Sixth GIFT Seminar*, Jaca, 2–7 June 1975. (Reprinted in Bell (2004), 52–62)

Bell, J. S. (1980). Atomic-cascade photons and quantum-mechanical nonlocality. *Comments on Atomic and Molecular Physics 9*, 121–26.

Bell, J. S. (1981). Bertlmann's socks and the nature of reality. *Journal of Physique, Colloque C"*, supp. au numero 3, Tome 42, 41–61.

Bell, J. S. (1986). EPR correlations and EPW distributions. In *New techniques and ideas in quantum measurement theory*. New York: Academy of Sciences. (Reprinted in Bell (2004), 196–200)

Bell, J. S. (1990). La nouvelle cuisine. In J. Sarlemijn & P. Kroes (Eds.), *Between science and technology*. Burlington: Elsevier. (Reprinted in Bell (2004), 232–248)

Bell, J. S. (2004). *Speakable and unspeakable in quantum mechanics*. Cambridge: Cambridge University Press.

Haag, R. (1992). *Local quantum physics*. Berlin: Springer.

Henson, J. (2013). Non-separability does not relieve the problem of Bell's theorem. *Foundations of Physics, 43*, 1008–1038.

Hofer-Szabó, G., & Vecsernyés, P. (2015a). On the concept of Bell's local causality in local classical and quantum theory. *Journal of Mathematical Physics, 56*, 032303.

Hofer-Szabó, G., & Vecsernyés, P. (2015b). Bell's local causality for philosophers (Synthese).

Mann, C., & Crease, R. (1988). John Bell, particle physicist (Interview). *Omni, 10/8*, 84–92.

Norsen, T. (2009). Local causality and completeness: Bell vs. Jarrett. *Foundations of Physics, 39*, 273.

Norsen, T. (2011). J.S. Bell's concept of local causality. *American Journal of Physics, 79*, 12.

Seevinck, M. P., & Uffink, J. (2011). Not throwing out the baby with the bathwater: Bell's condition of local causality mathematically 'sharp and clean'. In D. Dieks, W. J. Gonzalez, S. Hartmann, Th. Uebel, & M. Weber (Eds.), *Explanation, prediction, and confirmation* (The philosophy of science in a European perspective, Vol. 2, pp. 425–450). Dordrecht: Springer.

Pragmatists and Purists on CPT Invariance in Relativistic Quantum Field Theories

Jonathan Bain

1 Introduction

Pragmatist approaches to relativistic quantum field theories (RQFTs) trade mathematical rigor for the ability to derive predictions from realistic interacting theories. Examples include the Lagrangian approach found in most textbooks, and Weinberg's approach. *Purist* approaches to RQFTs trade the ability to formulate realistic interacting theories for mathematical rigor. Examples include the axiomatic and algebraic formalisms. Philosophers are split on whether foundational issues related to RQFTs should be framed within pragmatist or purist approaches. Wallace (2011), for instance, has argued that cutoff quantum field theory (CQFT), a particular pragmatist approach, has been successful at resolving the problems associated with renormalized perturbation theory, while axiomatic and algebraic quantum field theory (AQFT), which epitomize purist approaches, have not; and this indicates that CQFT is the correct framework for philosophy of QFT. Fraser (2011), on the other hand, argues that renormalization techniques indicate how CQFT and AQFT are empirically indistinguishable, and that AQFT is to be preferred for its mathematical rigor.

This essay probes this debate by viewing it through the lens of the CPT theorem. This theorem entails that the state of a physical system described by an RQFT must possess CPT invariance; i.e., invariance under the combined transformations of charge conjugation C, space inversion P, and time reflection T. There are both pragmatist and purist versions of this theorem (Bain 2013). While all versions apply unproblematically to non-interacting states, and some unrealistic interacting states,

J. Bain (✉)
Polytechnic School of Engineering, Department of Technology, Culture and Society, New York University, 6 Metrotech Center, Brooklyn, NY 11201, USA
e-mail: jon.bain@nyu.edu

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_17

extending them to realistic interacting states is problematic: For both pragmatists and purists, to do so requires confronting foundational problems. Greenberg (2002), however, claims that a violation of CPT invariance in an interacting RQFT, appropriately construed, entails a violation of Lorentz invariance. This claim is surprising not only since it purports to cover interacting theories in one fell swoop, but also because standard proofs of CPT invariance (both purist and pragmatist) require more than just the assumption of Lorentz invariance. Greenberg's claim has been influential in the physics literature since it suggests a test for violations of Lorentz invariance *via* experiments that measure CPT violation. Moreover, in apparently linking Lorentz invariance with CPT invariance, it suggests the latter is mysterious; in particular, some philosophers have wondered how the charge conjugation transformation C can arise from a purely spatiotemporal symmetry (Greaves 2010).

This essay analyzes Greenberg's claim in the context of the debate between pragmatists and purists. Section 2 reviews two formulations of the CPT theorem, one purist and the other pragmatist. Section 3 uses the problems these formulations face to inform a characterization of the distinction between pragmatism and purity based on the sense in which an RQFT can be said to exist. This distinction is then applied in Sect. 3 to a critique of Greenberg's claim. It will be seen that Greenberg's claim can be interpreted in either a purist or a pragmatist sense, and in either case, it fails to address the associated foundational problems.

2 Pragmatism Versus Purity on CPT Invariance

2.1 The Axiomatic CPT Theorem

The first example of a formulation of the CPT theorem I'd like to consider is the purist Wightman axiomatic approach (see, e.g., Streater and Wightman 1964).[1] The basic objects are vacuum expectation values of unordered products of fields, referred to as Wightman functions, $W^{(n)}(x_1, \ldots, x_n) \equiv \langle 0 | \phi(x_1) \ldots \phi(x_n) | 0 \rangle$, where $\phi(x)$ is a generic quantum field (technically defined as an operator-valued distribution), and $|0\rangle$ is its vacuum state. Wightman functions are required to satisfy a number of axioms, and it is the goal of this approach to construct models of these axioms that represent interacting RQFTs. For the purposes of deriving CPT invariance, the following three assumptions suffice.

(i) *Restricted Lorentz invariance* (RLI). The fields are invariant under the restricted Lorentz group $L_+^{\uparrow}$ (the subgroup of the Lorentz group connected to the identity that consists of Lorentz boosts but not parity or time reversal transformations).

[1] Another purist approach is the algebraic formalism which will not be discussed in this essay. CPT theorems have been proven in the algebraic approach by Borchers and Yngvason (2001) and Guido and Longo (1995). For a brief discussion of the latter, see Bain (2013).

(ii) *Spectrum Condition* (SC). The fields possess positive energy, in the sense that the spectrum of the momentum operator associated with $L_{+}^{\uparrow}$ is confined to the forward lightcone.

(i) and (ii) entail that Wightman functions can be extended to complex-analytic functions that are invariant under the proper complex Lorentz group. Moreover, the extended domain contains real points of analyticity referred to as Jost points.[2] The third assumption refers to these latter:

(iii) *Weak Local Commutativity* (WLC). At (or in the neighborhood of) a Jost point the fields satisfy $\langle 0 \left| \phi \left(x_1 \right) \ldots \phi \left(x_n \right) \right| 0 \rangle = i^K \langle 0 \left| \phi \left(x_n \right) \ldots \phi \left(x_1 \right) \right| 0 \rangle$, where K is the number of fermionic fields.

Jost (1957) showed that the conjunction of (i), (ii), (iii) entails the existence of an anti-unitary operator that combines the actions of C, P, and T transformations on fields, leaving them invariant (Streater and Wightman 1964, p. 150). The axiomatic CPT theorem thus states:

$$[(\text{RLI of fields}) \ \& \ \text{SC} \ \& \ \text{WLC}] \Rightarrow (\text{CPT invariance of fields})$$

This axiomatic understanding of CPT invariance faces what might be called the *Problem of Empirical Import*: No "realistic" interacting models of the Wightman axioms currently exist; i.e., no interacting models exist for theories (like QED and QCD) from which empirical predictions have been derived and confirmed. On the other hand, non-interacting models, and "unrealistic" interacting models of the Wightman axioms have been constructed (the latter are discussed by Fraser 2011, p. 127). This suggests that the axiomatic CPT theorem (currently) restricts CPT invariance to non-interacting, or unrealistic interacting RQFT states. This is problematic, since the evidence for CPT invariance in particular, and for the reliability of RQFTs in general, invariably comes from successful predictions made by realistic interacting RQFTs.

2.2 Weinberg's CPT Theorem

I'd now like to consider Weinberg's derivation of the CPT theorem as an example of a pragmatist approach (Weinberg 1995). The basic object of this approach is the S-matrix, which satisfies three assumptions:

(i) *Perturbation Theory*. The S-matrix is given by a power series expansion in time-ordered products of an interaction Hamiltonian density $\mathcal{H}_{int}(x)$:

[2] A Jost point $(x_1, \ldots, x_n)$ is a convex set of points that are spacelike separated from each other. In other words, the difference variables $\xi_i \equiv x_{i-1} - x_i$ satisfy $\left(\sum \lambda_j \xi_j\right)^2 < 0$, for $\lambda_j \geq 0, \sum \lambda_j > 0$ (Streater and Wightman 1964, p. 71).

$$S_{\beta\alpha} = \sum_{n=0}^{\infty} \frac{-i^n}{n!} \int \langle \beta \, | T \{ \mathcal{H}_{int}(x_1) \ldots \mathcal{H}_{int}(x_n) \} | \, \alpha \rangle \, d^4x_1 \ldots d^4x_n \qquad (1)$$

where $|\beta\rangle, |\alpha\rangle$ are asymptotic multi-particle states, and the time-ordered product $T\{\mathcal{H}_{int}(x_1) \ldots \mathcal{H}_{int}(x_n)\}$ orders the $\mathcal{H}_{int}(x_i)$ according to $t_1 > \cdots > t_n$.

(ii) *Lorentz Invariance*. The S-matrix is invariant under restricted Lorentz transformations.

(iii) *Cluster Decomposition* (CD). The S-matrix satisfies cluster decomposition (briefly, correlations between scattering experiments decrease to zero as their separation distance increases to space-like infinity).

Weinberg shows that a sufficient condition for CD to be compatible with (ii) is that $\mathcal{H}_{int}(x)$ be a functional of fields that satisfy RLI and local commutativity (i.e., the fields commute or anti-commute at spacelike separated distances), and that are linear combinations of Fock space creation and annihilation operators for non-interacting particle states. Weinberg (1995, p. 198) then argues that if these fields carry a conserved charge, then anti-particle states must be posited. CPT invariance of the full Hamiltonian density then follows from a consideration of how the relevant creation and annihilation operators transform under C, P, and T separately (1995, pp. 244–246). The CPT theorem thus takes the following form:

$$[(\text{RLI of } S\text{-matrix}) \ \& \ \text{CD} \ \& \ (\text{existence of conserved charges})]$$
$$\Rightarrow (\text{CPT invariance of } \mathcal{H}(x))$$

where $\mathcal{H}(x)$ is the full Hamiltonian density.

In Weinberg's approach, one might claim that CPT invariance is a property of both interacting and non-interacting states, insofar as the demonstration of CPT invariance of $\mathcal{H}(x)$ rests on CPT invariance of the creation and annihilation operators of non-interacting multi-particle states that transform, under the S-matrix, into interacting multi-particle states. However, lest one think that this is an improvement over the axiomatic understanding of CPT invariance, the rigor of this approach faces the following problems:

(a) Expression (1) assumes that multi-particle states at asymptotic times are non-interacting, and can be unitarily related to interacting states at finite times. This is made problematic by Haag's theorem, which indicates that, under reasonable assumptions, the Hilbert spaces for interacting and non-interacting states belong to unitarily inequivalent representations of the canonical (anti-)commutation relations, thus a unitary S-matrix operator that transforms non-interacting states into interacting states does not exist (Duncan 2012, pp. 359–370).

(b) For many of the types of interacting QFTs of interest, the terms in the power series (1) diverge at high energies. This is referred to as the *UV* (*ultra-violet*) *Problem*.

(c) For the types of interacting QFTs of interest, there is a consensus that the power series (1) does not converge. Call this the *Convergence Problem*.

A few qualifications are in order at this point. First, these problems are not unique to Weinberg's approach, but rather afflict pragmatist approaches in general. Second, some interacting QFTs of interest, quantum chromodynamics (QCD) for instance, do not suffer the *UV Problem*; it is generally thought that QCD has an ultra-violet fixed point (more on this in Sect. 3 below). Third, problem (a) is implicitly addressed in pragmatist approaches by employing renormalization. In order to further distinguish pragmatists from purists, it will help to review where in pragmatist approaches renormalization occurs. This will be done in Sect. 2.3 below, but before this discussion, a final concern should perhaps be addressed involving the extent to which the Wightman axiomatic CPT theorem differs from Weinberg's CPT theorem.

In many textbooks, one finds pragmatist proofs of the CPT theorem followed by the advice that if one seeks a more rigorous proof, one should consult the Wightman axiomatic approach (see, e.g., Weinberg 1995, p. 245; Duncan 2012, pp. 479–483). There is a limited sense in which such pragmatist appeals to purist proofs of the CPT theorem seem justified. In the cases of non-interacting theories, and in some unrealistic interacting theories, one can argue that purist approaches and pragmatist approaches are intertranslatable. In these cases, purists do not face the *Problem of Empirical Import*, and pragmatists do not face the *Convergence Problem* (in these cases, expressions like (1) converge, and, moreover, as explained below, problems (a) and (b) can be effectively addressed). When intertranslatability holds, a pragmatist might be excused for appealing to Jost's proof, for instance, to explain CPT invariance. However, even in such cases, it seems to me that if we take pragmatist and purist approaches literally, we should hesitate to "mix and match" proofs of CPT invariance. In particular, in Weinberg's approach, the basic objects are the *S*-matrix and particle states, and fields and field equations are purely instrumental devices. Weinberg makes it clear that fields are introduced only to guarantee that the *S*-matrix satisfies restricted Lorentz invariance and Cluster Decomposition.[3] Moreover, to accomplish this task, fields are introduced in a particular format; i.e., as linear combinations of creation and annihilation operators that act on a multiparticle Fock space. In the Wightman axiomatic approach, on the other hand, the basic objects, arguably, are fields which need not be expressible as linear combinations of Fock space creation and annihilation operators.[4] Thus, one reason an advocate of the Weinberg approach should be hesitant in adopting Jost's proof of CPT invariance is that the latter is more general than Weinberg's

[3]This is reflected in Weinberg's (1995, p. 198) view of the axiomatic assumption of local commutativity (i.e., fields at spacelike separated points commute): "The point taken here is that [local commutativity of fields] is needed for the Lorentz invariance of the *S*-matrix, without any ancillary assumptions about measurability or causality."

[4]In other words, a model of the Wightman axioms need not take the form of a Fock space representation of the canonical (anti-) commutation relations. Note that the basic objects of the Wightman approach are tempered distributions (i.e., Wightman functions), but Wightman's (1956) reconstruction theorem indicates that these can be interpreted as vacuum expectation values of unordered products of fields.

proof, and suggests the theory is about more than a pragmatist intends it to be about. Moreover, keeping purist and pragmatist proofs of the CPT theorem separate may be important when it comes to addressing the question of why the theorem fails for non-relativistic quantum field theories (and non-relativistic quantum mechanics in general). For instance, Lévy-Leblond (1967) appeals to Weinberg's proof to explain why the CPT theorem fails for axiomatic Galilei-invariant QFTs (i.e., Galilei-invariant QFTs formulated *via* a set of axioms similar to the Wightman axioms).[5] Given the conceptual differences between these approaches, and in particular, in those cases of physical interest in which intertranslatability fails, it seems more appropriate to frame an explanation of the failure of the CPT theorem in axiomatic non-relativistic QFTs in terms of Jost's axiomatic proof, as opposed to Weinberg's proof.

2.3 Pragmatism and the Renormalization Problem

Typical pragmatist approaches simplify (1) by reducing it to an expression that involves vacuum expectation values of time-ordered products of fields, $\langle 0 \left| T \left\{ \phi \left(x_1 \right), \ldots, \phi \left(x_n \right) \right\} \right| 0 \rangle$, referred to as τ-functions. One can distinguish between non-interacting and interacting τ-functions, depending on whether the fields are non-interacting or interacting (i.e., satisfy homogeneous or inhomogeneous field equations, respectively). The initial goal of pragmatist approaches is to reduce (1) to an expression that only involves non-interacting τ-functions (this subsequently facilitates the calculation of (1) *via* Feynman diagrams). This goal is achieved by the following:

(i) One first uses the LSZ reduction formula to relate S-matrix elements to interacting τ-functions (e.g., Duncan 2012, p. 286). This formula comes in many flavors, one per type of field. For instance, the LSZ formula for a scalar field of mass m is given by:

$$
\begin{aligned}
&{}_{out}\left\langle \mathbf{p}_1, \ldots, \mathbf{p}_n \middle| \mathbf{k}_1, \ldots, \mathbf{k}_\ell \right\rangle_{in} \\
&= \left(i/\sqrt{Z} \right)^{n+\ell} \int d^4x_1 \ldots d^4y_\ell e^{-ip_i x_i + ik_j y_j} \prod_i \left(\partial_{x_i}^2 + m^2 \right) \prod_j \left(\partial_{y_j}^2 + m^2 \right) \\
&\quad \times \langle 0 \left| T \left\{ \varphi \left(x_1 \right) \ldots \varphi \left(x_n \right) \varphi \left(y_1 \right) \ldots \varphi \left(y_\ell \right) \right\} \right| 0 \rangle .
\end{aligned}
\tag{2}
$$

[5] Lévy-Leblond (1967, p. 165) explains the failure of the CPT theorem in GQFTs as due to the fact that GQFTs do not satisfy local commutativity: "This situation [i.e., the GQFT case] is to be contrasted with the relativistic case where the requirements of local commutativity on a free field ... impose both the existence of a TCP [i.e., CPT] operation ... and the spin-statistics relation, as has been shown in a very illuminating way, for this free-field case, by Weinberg ..."

The left-hand side of (2) represents an S-matrix element for ℓ incoming particles with momenta k_i and n outgoing particles with momenta p_i. The right-hand side indicates how this can be calculated in terms of an interacting τ-function, where $\varphi(x)$ is an interacting field (i.e., a solution to the inhomogeneous Klein-Gordon equation).

(ii) One then assumes a perturbative split of the Hamiltonian, $H = H_0 + H_{int}$, into a non-perturbed piece H_0 and a piece H_{int} encoding small perturbations away from H_0.[6] The following Gell-Mann/Low formula then relates interacting τ-functions to non-interacting τ-functions (e.g., Duncan 2012, p. 246):

$$\langle 0 \left| T \left\{ \varphi \left(x_1 \right) \ldots \varphi \left(x_n \right) \right\} \right| 0 \rangle = \frac{\left\langle 0 \left| T \left\{ \phi_I \left(x_1 \right) \ldots \phi_I \left(x_n \right) e^{-i \int H_I dt} \right\} \right| 0 \right\rangle}{\left\langle 0 \left| T \left\{ e^{-i \int H_I dt} \right\} \right| 0 \right\rangle}. \tag{3}$$

In (3), $\varphi(x)$ is an interacting field, $\phi_I(x)$ is a non-interacting field in the interaction picture, and $H_I \equiv e^{iH_0} H_{int} e^{-iH_0}$ is the interaction picture representation of H_{int}.[7]

In the LSZ formula (2), Z is a renormalization constant. Its purpose is to relate the interacting field $\varphi(x)$ to non-interacting fields $\phi_{in}(x), \phi_{out}(x)$ at asymptotic times. One assumes,

$$\langle \beta \left| \varphi(x) \right| \alpha \rangle \underset{t \to -\infty}{\longrightarrow} \sqrt{Z} \langle \beta \left| \phi_{in}(x) \right| \alpha \rangle , \langle \beta \left| \varphi(x) \right| \alpha \rangle \underset{t \to +\infty}{\longrightarrow} \sqrt{Z} \langle \beta \left| \varphi_{out}(x) \right| \alpha \rangle \tag{4}$$

where $|\beta\rangle , |\alpha\rangle$ are non-interacting multi-particle states. This assumption may be motivated by considering the action of a non-interacting asymptotic field on the vacuum with respect to a single-particle state (Duncan 2012, p. 282). If $|\mathbf{k}\rangle$ is a normalized single-particle state, then $\langle \mathbf{k} \left| \phi_{in}(x) \right| 0 \rangle = 1$. An interacting field $\varphi(x)$ cannot, in general, be decomposed into creation and annihilation operators, thus one sets $\langle \mathbf{k} \left| \varphi(x) \right| 0 \rangle = \sqrt{Z}$, for some constant Z. (4) may be considered a generalization of this. Formally, the constant Z can be removed from the LSZ formula by replacing the "bare" interacting field with a renormalized interacting field defined by $\varphi_r(x) \equiv Z^{-1/2} \varphi(x)$. This assignment guarantees that the renormalized interacting field behaves like the non-interacting field with respect to single-particle states; namely, $\langle \mathbf{k} \left| \varphi_r(0) \right| 0 \rangle = 1$.

Renormalization also enters into the derivation of the Gell-Man/Low formula (3). In particular, (3) assumes $H_0 |0\rangle = 0 = H |0\rangle$. The first equality entails $|0\rangle$ is the vacuum state of the non-interacting fields. Since H is a functional of interacting fields which cannot, in general, be decomposed into creation and annihilation operators, the second equality is typically not guaranteed. To enforce it, one defines a renormalized Hamiltonian $H_r \equiv H - \Delta$. This corresponds to renormalizing the

[6] The Hamiltonian is related to the Hamiltonian density by $H(t) = \int d^3\mathbf{x} \mathcal{H} (\mathbf{x}, t)$.

[7] $\phi_I(x)$ is defined by $\phi_I (\mathbf{x}, t) \equiv e^{iH_0(t-t_0)} \phi (\mathbf{x}, t_0) e^{-iH_0(t-t_0)}$, where $\phi(\mathbf{x}, t_0)$ is a non-interacting field at time t_0.

mass that appears in H. If this is given by m_B (the "bare" mass), and the shift corresponding to Δ is given by δm, then the renormalized mass m_r (the "physical" mass) is given by ${m_r}^2 \equiv {m_B}^2 + \delta m^2$

In these examples, renormalization is imposed to force the interacting theory to behave like the non-interacting theory, as far as the vacuum and single-particle states are concerned. This solves Problem (a) in the following sense: The renormalized field and the renormalized Hamiltonian are not self-adjoint operators (for typical interacting theories, the constant Z and the mass shift δm are infinite). This entails, for instance, that H_r does not implement unitary time translations, contrary to one of the assumptions of Haag's theorem (Fraser 2009, p. 547). Whether this constitutes an *adequate* solution to Problem (a) will depend on one's mathematical proclivities. The fact that renormalized parameters are, typically, infinite may upset purists. For such purists, the renormalization procedure simply replaces Problem (a) with another problem, call it the *Renormalization Problem*.

Note that renormalization is independent of perturbation theory as evidenced by its appearance in the non-perturbative LSZ formula. Thus the *Renormalization Problem* is independent of the *UV* and *Convergence Problems*.[8] At this point, it will be instructive to review how renormalization group (RG) techniques address these pragmatist problems. Wallace (2011) argues that such techniques underwrite heuristic (i.e., pragmatist) approaches, whereas Fraser (2011) claims they support rigorous (i.e., purist) approaches. The next section addresses this issue, as well as the general concern of how best to distinguish purity from pragmatism.

3 Distinguishing Purity from Pragmatism

The goal of the RG approach to renormalization is to determine how a theory's low-energy degrees of freedom depend on its high-energy degrees of freedom. Towards this end, the coupling constants g that appear in the interaction Hamiltonian (or Lagrangian) density, are defined as functions $g(\Lambda(\mu))$ of a scale-dependent cutoff $\Lambda(\mu)$. Changing the scale (by integrating out high-energy degrees of freedom with respect to Λ) generates a flow in the theory's parameter space. Couplings can then be characterized by how they behave as the scale is lowered: relevant couplings increase, irrelevant couplings decrease, and marginal couplings remain constant. One can show that, for a $(3+1)$-dim weakly coupled theory, there are a finite number of relevant and marginal couplings, and any irrelevant couplings are suppressed at a given energy scale μ by powers of μ/Λ(e.g., Duncan 2012, pp. 652–660).[9] In such a theory, the low-energy degrees of freedom depend on the

[8] As Weinberg (1995, p. 441) states, "...the renormalization of masses and fields has nothing directly to do with the presence of infinities, and would be necessary even in a theory in which all momentum space integrals were convergent."

[9] An important exception to this is QCD, which is not weakly coupled.

high-energy degrees of freedom through a finite number of parameters (the relevant and marginal couplings), and while the theory may still contain parameters that become infinite at high energies (the irrelevant couplings), it is still predictive in the sense that its predictions will be finite if constrained to a given scale. At this scale, the theory is *effectively* renormalizable insofar as any irrelevant couplings it may possess cannot be experimentally detected. With respect to Sect. 2.3's discussion of renormalization, an effectively renormalizable interacting theory requires a finite number of parameters to empirically imitate the behavior of the corresponding non-interacting theory, and these parameters, as functions of a finite cutoff, are finite.

In this effective field theory approach, the *Renormalization Problem* is addressed by adopting effective renormalizability, and the *UV Problem* is addressed by using the cutoff Λ to regulate divergent terms in expressions like (1) in Sect. 2.2. The cutoff serves to freeze out the high energy degrees of freedom of the theory, and one then adopts an agnostic attitude about what happens at energy scales above Λ. According to Wallace,

> This, in essence, is how modern particle physics deals with the renormalization problem: it is taken to presage an ultimate failure of quantum field theory at some short lengthscale, and once the bare existence of that failure is appreciated, the whole of renormalization theory becomes unproblematic, and indeed predictively powerful in its own right. (Wallace 2011, p. 119.)

While this appeal to RG techniques allows a pragmatist to address the *Renormalization* and *UV Problems*, the *Convergence Problem* still remains (with the qualifications noted at the end of Sect. 2.2). Moreover, Fraser (2011) suggests that RG techniques support purity, as opposed to pragmatism. In particular, the RG flow of the type of theory described above indicates an underdetermination of the theory's high-energy content by low-energy experiments. The latter fix the values of the theory's finite relevant and marginal couplings at the experimental energy scale, but fail to fix the values of the theory's irrelevant couplings. These latter determine how the theory behaves at high-energies. This implies that the successful predictions made by a realistic interacting RQFT (of this type) fail to determine the form it takes at high-energies. This suggests to Fraser that axiomatic and algebraic RQFT (AQFT) on the one hand, and Wallace's "cutoff" QFT (CQFT) on the other, are empirically indistinguishable at the energy scales currently probed by experiments:

> The upshot of the application of RG methods is that a range of Lagrangians at short distance scales each yield approximately the same predictions for relatively low energies. ... This lends support to the claim that the theoretical framework of QFT is underdetermined by the empirical evidence. AQFT and [CQFT] should be viewed as alternative theoretical frameworks for QFT which approximately agree in their empirical predictions. (Naturally, subject to the qualification that the construction of models of AQFT is still in progress). (Fraser 2011, p. 135.)

The qualification at the end of this quote is important. It acknowledges that the purist's *Problem of Empirical Import* is a potential obstruction to the claim that RG underdetermination holds between AQFT and CQFT. This obstruction takes the form of the question of whether there are AQFTs that can be "RG-related" to the appropriate low-energy experiments (as CQFTs can be).

These considerations suggest that an appeal to RG techniques is not decisive in adjudicating between pragmatists and purists. Both pragmatists and purists can make such an appeal, and such appeals fail to completely address foundational issues: the *Convergence Problem* remains for the RG pragmatist (with the requisite qualifications), and the *Problem of Empirical Import* remains for the RG purist. The discussion in Sect. 2.3 also indicates that an appeal to perturbation theory won't help either. On the one hand, pragmatists can employ non-perturbative techniques (the LSZ formula, for example; and lattice techniques in theories like QCD that are not weakly coupled). On the other hand, purists can employ perturbative techniques, as evidenced by "perturbative" AQFT which seeks to combine techniques from causal perturbation theory with AQFT (see, e.g., the review in Summers 2012, pp. 45–48).

The diversity of these methods allowed by both pragmatists and purists also suggests that a general appeal to mathematical rigor may not be enough to make the distinction as clear as it could be. Note first that the distinction between non-perturbative and perturbative methods does not necessarily map onto a distinction between rigorous and non-rigorous methods. In particular, the use of perturbative methods need not signal a relaxation of rigor. For instance, causal perturbation theory has been viewed by its advocates as providing a rigorous mathematical foundation for perturbative techniques, and these advocates include both purists and pragmatists.[10] Arguably, the lack of rigor that purists have traditionally associated with pragmatists' use of perturbation theory ultimately manifests itself in the pragmatists' *Convergence Problem*.

Thus what remains to distinguish pragmatists from purists are the *Convergence Problem* for the former (with the requisite qualifications), and the *Problem of Empirical Import* for the latter. These problems are concerned with the sense in which realistic interacting RQFTs can be said to exist. Call this basic foundational concern common to both purity and pragmatism, the *Existence Problem*. As Bouatta and Butterfield (2014, p. 16) suggest, this problem can be addressed in a number of ways. Purists, perhaps, can be essentially characterized by their demand for a strong notion of existence; namely, existence of a model of an appropriate set of axioms. Pragmatists, perhaps, can be essentially characterized by their adoption of a weaker notion of existence. One might require existence of a theory to entail the convergence of power series expansions like (1) (in which case interacting QED probably does not exist, whereas interacting QCD probably does). Alternatively, pragmatists might settle for existence defined in terms of renormalizability (in which case both interacting QED and QCD exist), or in terms of the existence of

[10]Causal perturbation theory consists of both a regularization scheme to address UV divergences in power series expansions, and an axiomatic scheme underwriting such expansions. These schemes can be separated; in particular, the regularization scheme can be adopted by pragmatists independently of the axiomatic scheme (Helling 2012; Falk et al. 2010). Conversely, the axiomatic scheme can be adopted by purists to extend purist axiomatic systems to include perturbative techniques (Brunetti and Fredenhagen 2000).

a UV fixed point in an RG flow (in which case interacting QED does not exist, but asymptotically free and/or safe theories like interacting QCD do, as well as conformally invariant theories).

Thus, distinguishing purity from pragmatism on the basis of the *Existence Problem* addresses the fact that both purists and pragmatists make use of similar methods, perturbative and non-perturbative, as well as the concern that mathematical rigor may be in the eye of the beholder. Moreover, it addresses the concern that the types of interactions described by realistic interacting RQFTs differ in essential ways. The next section will put this distinction to work.

4 Greenberg on Relativity and CPT Invariance

Greenberg (2002, p. 1) claims: "If CPT invariance is violated in an interacting quantum field theory, then that theory also violates Lorentz invariance." This claim is both influential and puzzling. In the physics literature it is cited for statements like the following:

> Note that Lorentz violation does not imply CPT violation for local EFTs, while CPT violation does imply Lorentz violation in local EFTs. (Liberati 2013, p. 12.)

> In all proofs of the CPT theorem Lorentz symmetry is the basic hypothesis, and indeed a theorem states that if CPT symmetry is violated then Lorentz symmetry must be violated, too . . . (Sozzi 2008, p. 198.)

> In realistic field theories, CPT violation is always accompanied by Lorentz violation, but not vice versa. (Berger 2011, p. 180.)

While Greenberg is not directly cited in the philosophy literature, one does find the following statements:

> . . . the CPT theorem . . . says that violations of CPT symmetry imply violations of Lorentz invariance, but not vice versa. (Hagar 2009, p. 261.)

> How can it come about that one symmetry (e.g., Lorentz invariance) entails another (e.g., CPT) *at all*? (Greaves 2010, p. 28)

> The CPT theorem says that any (restricted) Lorentz invariant quantum field theory must also be invariant under the combined operation of [CPT]. (Arntzenius 2011, p. 633.)

Greenberg's claim is puzzling for two reasons. First, in both the purist and pragmatist proofs of the CPT theorem reviewed in Sect. 2, more than just Lorentz invariance was needed to derive CPT invariance. Second, both proofs showed that, given appropriate assumptions, CPT invariance holds for non-interacting fields, and certain unrealistic interacting fields. In order to extend the proofs to realistic interacting fields, both the purist and the pragmatist need to confront Sect. 3's *Existence Problem*. This problem intimately depends on the type of interaction, and this suggests that a demonstration of CPT invariance for realistic interacting fields may have to be done on a case by case basis. Thus, on the surface, Greenberg's claim seems to both simplify the assumptions needed to derive CPT invariance, and address the issue of the extent of its applicability in one fell swoop.

Greenberg begins with the following assertions:

> To calculate the S matrix, we need τ functions, or similar functions, such as retarded or advanced products (r functions or a functions). We require covariance of a quantum field theory both in and out of cone as the condition for Lorentz invariance of the theory; thus both Wightman functions and the τ (or r or a) functions must be covariant for the theory to be Lorentz invariant. (Greenberg 2002, p. 1.)

He then provides the following expression for an n-point τ-function:

$$\tau^{(n)}\left(x_1,\ldots,x_n\right) \equiv \sum\nolimits_p \theta\left(t_{p_1}-t_{p_2}\right)\ldots\theta\left(t_{p_{n-1}}-t_{p_n}\right) W^{(n)}\left(x_{p_1},\ldots,x_{p_n}\right) \tag{5}$$

where the product of Heaviside functions $\theta\left(t_{p_1}-t_{p_2}\right)\ldots\theta\left(t_{p_{n-1}}-t_{p_n}\right)$ enforces the time ordering $t_{p_1} > \cdots > t_{p_n}$ on the Wightman function $W^{(n)}$, and the sum is over all permutations of the indices. Greenberg now argues that, at a Jost point, restricted Lorentz invariance of $\tau^{(n)}$ entails that $W^{(n)}$ satisfies weak local commutativity (WLC).[11] Thus if we require Wightman functions to satisfy RLI and SC, then a violation of CPT invariance of Wightman functions entails a violation of RLI of τ-functions. Schematically,

$$\begin{aligned}\left(\text{RLI of } \tau^{(n)} \text{at Jost points}\right) &\Rightarrow \left(\text{WLC of } W^{(n)} \text{at Jost points}\right)\\ &\Rightarrow \left(\text{CPT invariance of } W^{(n)} \text{that satisfy RLI and SC}\right)\end{aligned} \tag{6}$$

where the second entailment follows from the axiomatic proof of CPT invariance (Sect. 2.1). Given the assumption that a theory is RLI only if both its Wightman and τ-functions are RLI, Greenberg concludes that a violation of CPT invariance of a theory's Wightman functions entails the theory does not satisfy RLI. No mention of an *interacting* theory has occurred at this point. However Greenberg (2002, p. 1) now states: "[t]his argument does not apply to a non-interacting theory for which τ functions need not be considered". This suggests the view that τ functions are a necessary ingredient in interacting QFTs, but not in non-interacting QFTs. The complete argument may thus be schematically represented by the following:

I. RLI violation of τ-functions entails RLI violation of the corresponding interacting QFT.
II. CPT violation of Wightman functions entails RLI violation of the corresponding τ-functions.
$\therefore$ Therefore, CPT violation of Wightman functions entails RLI violation of the corresponding interacting QFT.

Dütsch and Gracia-Bondía (2012, p. 429) observe that Greenberg's argument depends on the assumption that expression (5) exists for realistic interacting theories. This of course faces the purist's *Problem of Empirical Import*. It appears

[11] The proof of this claim rests on the fact that it is always possible to choose two Lorentz transformations that time-order a Jost point $(x_1, \ldots, x_n)$ in opposite ways.

explicitly in Premise II, which relies on the axiomatic proof of CPT invariance. Recall that this proof assumes (among other things) that Wightman functions satisfy the Spectrum Condition, which is essential to establish that complex extensions of Wightman functions are analytic. Dütsch and Gracia-Bondía (2012, p. 429) then observe: "...to the best of our knowledge, for non-trivial realistic models one cannot ascertain analyticity of Wightman-like functions; hence the argument *a la* Jost in [Greenberg 2002] flounders." They conclude with the following remarks:

> While the assertion that PCT conservation holds for everyday interacting relativistic theories remains plausible, to the question whether it has been proven at the required level of rigour, the clear and present answer is: only for a class of models ... and for none by Greenberg's argument. (Dütsch and Gracia-Bondía 2012, p. 429.)

Thus, as a *purist* attempt to extend CPT invariance to realistic interacting RQFTs, Greenberg's argument fails, to the extent that it fails to address the obstacle to extending the standard axiomatic proof of CPT invariance to realistic interacting RQFTs; namely, the *Problem of Empirical Import*. One way to express this failure is the observation that simply replacing Wightman functions with τ-functions does not automatically convert a theory that satisfies CPT invariance according to the axiomatic proof into a realistic interacting theory.

Does Greenberg's argument fair any better as a *pragmatist* attempt to extend CPT invariance to realistic interacting RQFTs? It appears that pragmatists have good reason to reject both Premises I and II. Consider, first, how a pragmatist might view Premise I. In Weinberg's approach, for instance, we have the following implications (Weinberg 1995, pp. 144–145)s:

$$(\mathcal{H}_{int}(x) \text{ is RLI and commutes at spacelike separations}) \Rightarrow (\tau\text{-functions of } \mathcal{H}_{int}(x) \text{are RLI})$$
$$\Rightarrow (\text{RLI of } S\text{-matrix})$$

where $\mathcal{H}_{int}(x)$ is the theory's interaction Hamiltonian density.[12] Thus if an RQFT is identified with its S-matrix, then a violation of RLI of its τ-functions does not necessarily entail a violation of RLI of the theory. This immediately blocks Greenberg's argument without further discussion. On the other hand, if an RQFT is identified with its Hamiltonian density, then a violation of RLI of its τ-functions entails either the theory violates RLI, or it is nonlocal (in the sense that its Hamiltonian density does not commute at spacelike separations). Thus a way is still open for this type of pragmatist to avoid Greenberg's argument, too.[13]

[12]The first entailment is based on the fact that the time-ordering of two points is RLI unless the points are spacelike separated. Thus if a field is RLI, then so are time-ordered products of it, except when it is evaluated at spacelike separated points. But if the field commutes when it is evaluated at spacelike separated points, then time-ordering will not violate RLI even at such points. This also holds for sums of products of fields, and hence for $\mathcal{H}_{int}(x)$. The second entailment follows since if time-ordered products of $\mathcal{H}_{int}(x)$ are RLI, then so is the S-matrix in the form (1), since all other quantities in (1) are manifestly RLI.

[13]Chaichian et al. (2011, p. 178) provide examples of non-local interaction Hamiltonian densities that are restricted Lorentz invariant and violate CPT invariance (thanks to a referee for pointing this out).

Two observations perhaps should be made at this point. First, note that one can adopt either of these options (i.e., identifying an RQFT with its *S*-matrix or with its Hamiltonian density) and still be faced with the pragmatist's *Existence Problem*. One is still faced with the question of whether a given *S*-matrix is well-defined in any of the pragmatist senses listed in Sect. 3, or if a given Hamiltonian density entails a well-defined *S*-matrix in these senses. Second, one might argue that the type of RQFTs of interest should be local (in the sense that their Hamiltonian densities commute at spacelike separations), and/or should be such that the condition of RLI of time-ordered products of their Hamiltonian densities is both necessary and sufficient for RLI of their *S*-matrix. But more must be said on both points for Greenberg's argument to gain initial traction for pragmatists.[14]

With respect to Premise II, pragmatists can justify the existence of realistic interacting τ-functions, not by providing provisos concerning the possibility of constructing realistic interacting models of a set of axioms, but rather by employing the Gell-Mann/Low formula (3). However, this confronts them with the *Existence Problem*. In particular, the Gell-Mann/Low formula requires a perturbative power series expansion of relevant quantities, and this expansion, even after it has been regularized and renormalized, fails to converge for the theories of interest. Thus, with respect to Premise II, Greenberg's argument is on the same shaky foundations for pragmatists as it is for purists. This problem makes its explicit appearance for a pragmatist in the second entailment in Greenberg's derivation (6) of Premise II. The technical difficulty in this case is that realistic interacting τ-functions obtained from the Gell-Mann/Low formula do not satisfy the Spectrum Condition.[15]

The upshot of this discussion is that, considered as either a purist or a pragmatist attempt to extend CPT invariance to realistic interacting fields, Greenberg's claim faces the *Existence Problem*. While his demonstration is insightful in uncovering connections between Lorentz invariance and CPT invariance in abstract objects like τ-functions and Wightman functions, both purists and pragmatists should be hesitant in extending these observations to concrete things like realistic interacting RQFTs.

[14]Here is another concern about the feasibility of Premise I in pragmatist approaches. If an interacting RQFT is in the business of calculating *S*-matrix elements, then τ-functions play an important role, as the discussion of the LSZ and Gell-Mann/Low formulas indicated, and this seems to make Premise I initially plausible. However, if there are other methods for calculating *S*-matrix elements that do not rely on τ-functions, and, moreover, if there are other testable predictions of RQFTs that can be derived without the use of τ-functions, then Premise I will again loose traction with pragmatists.

[15]For the purist, this problem manifested itself in the fact that currently there are no examples of realistic interacting τ-functions in the form of well-defined analytic functions. For the pragmatist who allows τ-functions to take the form of divergent power series expansions obtained *via* the Gell-Mann/Low formula, the problem is that such expressions do not satisfy the Spectrum Condition (in the sense that the fields that occur in them do not satisfy the Spectrum Condition).

5 Conclusion

This essay has used the debate between purists and pragmatists to critically examine Greenberg's (2002) claim that a violation of CPT invariance in an interacting RQFT entails a violation of Lorentz invariance. Section 2 revealed the extent to which purist and pragmatist versions of the CPT theorem extend to realistic interacting RQFTs. In both cases, this extent is constrained by what Sect. 3 called the *Existence Problem*; namely, the problem of articulating an appropriate notion of existence for a QFT, and then demonstrating that this notion holds for realistic interacting RQFTs. Purists can be characterized by their adoption a notion of existence that requires the existence of a model of an appropriate set of axioms, and the *Existence Problem* then becomes the task of constructing such a model for realistic interacting RQFTs. Pragmatists can be characterized by their adoption of a weaker notion of existence (convergence, renormalizability, existence of a UV fixed point, *etc.*), and the *Existence Problem* then becomes the task of demonstrating that their preferred notion holds for the types of realistic interacting RQFTs of interest. Greenberg's claim was shown in Sect. 4 to suffer from a failure to address the *Existence Problem*, in either its purist or its pragmatist form.

References

Arntzenius, F. (2011). The CPT theorem. In C. Callender (Ed.), *The Oxford handbook of philosophy of time* (pp. 633–646). Oxford: Oxford University Press.

Bain, J. (2013). CPT invariance, the spin-statistics connection, and the ontology of relativistic quantum field theories. *Erkenntnis, 78*, 797–821.

Berger, M. (2011). Lorentz violation in top-quark production and decay. In V. Kostelecky' (Ed.), *Proceedings of the 5th meeting on CPT and Lorentz Symmetry* (pp. 179–183). Singapore: World Scientific.

Borchers, H., & Yngvason, J. (2001). On the PCT-theorem in the theory of local observables. In R. Longo (Ed.), *Mathematical physics in mathematics and physics: Quantum and operator algebraic aspects* (pp. 39–64). Providence: American Mathematical Society. Available online as arXiv:math-ph/0012020v1.

Bouatta, N., & Butterfield, J. (2014). On emergence in gauge theories at the 't Hooft limit. *European Journal for Philosophy of Science*. doi:10.1007/s13194-014-0098-1. Preprint. Available at http://philsci-archive.pitt.edu/id/eprint/9288.

Brunetti, R., & Fredenhagen, K. (2000). Microlocal analysis and interacting quantum field theories: Renormalization on physical backgrounds. *Communications in Mathematical Physics, 208*, 623–661.

Chaichian, M., Dolgov, A., Novikov, V., & Tureanu, A. (2011). CPT violation does not lead to violation of lorentz invariance and vice versa. *Physics Letters, B699*, 177–180.

Duncan, A. (2012). *The conceptual framework of quantum field theory*. Oxford: Oxford University Press.

Dütsch, M., & Gracia-Bondía, J. (2012). On the assertion that PCT violation implies Lorentz non-invariance. *Physics Letters, B711*, 428–433.

Falk, S., Häußling, R., & Scheck, F. (2010). Renormalization in quantum field theory: An improved rigorous method. *Journal of Physics, A43*, 035401.

Fraser, D. (2009). Quantum field theory: Underdetermination, inconsistency, and idealization. *Philosophy of Science, 76*, 536–567.

Fraser, D. (2011). How to take particle physics seriously: A further defense of axiomatic quantum field theory. *Studies in History and Philosophy of Modern Physics, 42*, 126–135.

Greaves, H. (2010). Towards a geometrical understanding of the CPT theorem. *British Journal for the Philosophy of Science, 61*, 27–50.

Greenberg, O. (2002). CPT violation implies violation of Lorentz invariance. *Physical Review Letters, 89*, 231602.

Guido, D., & Longo, R. (1995). An algebraic spin and statistics theorem. *Communications in Mathematical Physics, 172*, 517–533.

Hagar, A. (2009). Minimal length in quantum gravity and the fate of Lorentz invariance. *Studies in History and Philosophy of Modern Physics, 40*, 259–267.

Helling, R. (2012). *How I learned to stop worrying and love QFT* (preprint), arXiv:1201.2714v2 [math-ph].

Jost, R. (1957). Eine Bemerkung zum CTP theorem. *Helvetica Physica Acta, 30*, 409–416.

Lévy-Leblond, J.-M. (1967). Galilean quantum field theories and a ghostless Lee model. *Communications in Mathematical Physics, 4*, 157–176.

Liberati, S. (2013). Tests of Lorentz invariance: A 2013 update. *Classical and Quantum Gravity, 30*, 133001.

Sozzi, M. (2008). *Discrete symmetries and CP violation*. Oxford: Oxford University Press.

Streater, R., & Wightman, A. (1964). *PCT, spin and statistics, and all that*. Princeton: Princeton University Press.

Summers, S. (2012). *A perspective on constructive quantum field theory* (preprint), arXiv:1203.3991v1 [math-ph].

Wallace, D. (2011). Taking particle physics seriously: A critique of the algebraic approach to quantum field theory. *Studies in History and Philosophy of Modern Physics, 42*, 116–125.

Weinberg, S. (1995). *The quantum theory of fields* (Vol. 1). Cambridge: Cambridge University Press.

Wightman, A. (1956). Quantum field theory in terms of vacuum expectation values. *Physical Review, 101*, 860–866.

Explanation in Quantum Chemistry

Carsten Seck

1 Quantum Chemistry

The historical roots of quantum chemistry - a branch of chemistry that utilizes quantum mechanics to explain chemical phenomena can be traced back to the emergence and rise of physical chemistry in the nineteenth and early twentieth century. The more general focus of the latter is the application of physical methods and theories like, for instance, spectroscopy and thermodynamics to chemistry. Although the evolution of physical chemistry in general and quantum chemistry in particular was essentially boosted by the work of Gilbert Lewis, Fritz London, Walter Heitler and Linus Pauling (cf. Gavroglu and Simoes 2012), I would like to begin with a famous diagnosis of an likewise influential and 'hegemonic' perspective towards quantum mechanics and its application to molecular scale range phenomena, made by a full-hearted theoretical physicist. Paul Dirac enthusiastically stated in his paper on 'Quantum Mechanics of Many-Electron Systems' (Dirac 1929, p. 714):

> The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.

Dealing with the inter-theoretic relation of physics and chemistry, Dirac significantly frames the subsequent research program of quantum chemistry to develop approximate practical methods of applying quantum mechanics. The decisive point here is that even genuine chemical explanations of, say, the properties of molecular species are ultimately gaining their explanatory status through 'necessary' physical

C. Seck (✉)
Department of Philosophy, University of Bonn, Am Hof 1, Bonn D-53113, Germany
e-mail: carsten.seck@uni-bonn.de

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_18

components.[1] To this effect, Dirac is frequently regarded as having reinforced a reductionist strand in chemistry (cf. Hendry 2012, p. 369) that is driven by the claim of deriving everything from first principles, or in other words *ab initio*. Eric Scerri recently described this ideal using the very illustrative analogy that the "epitome of the ab initio approach is something like Euclidean geometry, where one begins with a number of axioms and one derives everything from the starting point without any recourse whatsoever to empirical data" (Scerri 2004, p. 94).

In this sense an explanation of the dynamics of molecules, fluids, solutions and liquid crystals would ultimately be a physical, a quantum mechanical task: higher level species are to be exhaustively reduced to interacting particles of fundamental physics. To espouse such *Euclidean ideal* in contemporary chemistry, we surely have to consider physical chemistry that investigates physicochemical phenomena using particularly techniques from atomic and molecular physics. A very exciting cutting-edge branch of physical chemistry is the so-called computational quantum chemistry. This research area explicitly tends to put *ab initio* calculations to work "in which no experimental data whatsoever are admitted into the computation" aiming to simply "calculate the energy of a molecule, a bond angle, a dipole moment, or rate of reaction from the first principles of quantum mechanics" (Scerri and McIntyre 1997, p. 216).

Certainly, the problem of a reduction of chemistry to physics is an enduring theme in the philosophy of chemistry. Typically, philosophers of chemistry are doubtful concerning a full reduction of chemistry to the reign of quantum mechanics (cf. Hendry 2012). As a matter of fact, I fully endorse the respective arguments in Scerri (1997, 2004), Gavroglu (1997), Hendry and Needham (2007) and Needham (2010). These arguments are usually arranged around the question of whether or not specific chemical concepts and findings such as 'valence', 'bond' or the 'length of periods in the periodic table' can be adequately reconstructed and/or derived from quantum mechanics.

In the following sections of this quite programmatic paper I am neither going to add a further argument of this kind nor do I pretend to give a sufficient coverage of any of the arising philosophical issues.[2] Rather, I am going to take into account the practice of *ab initio* molecular dynamics, which would at first sight clearly be an eligible candidate for implementing a kind of a reductionist program. I suggest that a reductionist philosopher of chemistry is not going to find (even in the practice of this cutting-edge branch of computational chemistry) any adequate mounting point for a reductionist position.

[1] I leave open the question of whether Dirac himself values the underlying physical laws as *sufficient* to flesh out the explanatory tasks of 'the whole chemistry' as well.

[2] An exhaustive discussion would certainly need to address the paradigmatic issues and literature centered around reduction, emergence, semi-classical modelling etc. in more detail.

2 *Ab Initio* Molecular Dynamics

The so-called *ab initio* molecular dynamics explores the structure and dynamics of complex molecular many-body systems through computer simulations. Quantum chemists coined the term 'ab initio' to indicate an algorithmic simulation of chemical reactions of molecules in a kind of a 'virtual laboratory', which is – as frequently stated by quantum chemists – governed solely by the basic laws of physics. In order to get such simulations to work, a specific software package is needed. The developers of *Gaussian 09*, the latest version of a software environment that is used by many chemists, chemical engineers, biochemists, physicists and other scientists around the world, endorse a kind of the aforementioned *Euclidean ideal* quite enthusiastically (Gaussian 2014):

> Starting from the fundamental laws of quantum mechanics, Gaussian 09 predicts the energies, molecular structures, vibrational frequencies and molecular properties of molecules and reactions in a wide variety of chemical environments.

Taking such an attitude seriously we could, at first sight, conceive of molecular dynamics as an instrument of universal reduction. Scientists who are faced with some problem must only identify the involved molecular constellation and enter the respective parameters (that are organized in comprehensive databases) into the number crunching methods of Gaussian. Then, after a complex sequence of program instructions is performed, they could ultimately obtain explanations and predictions which would indeed reveal impressive insights in how objects of the sciences are to be decomposed to the time-evolution of interacting electrons and nuclei as described in the framework of quantum mechanics. In order to scrutinize this salacious claim, I shall take a closer look at what is really at stake in *ab initio* molecular dynamics.

As it turns out, the models underlying simulations of molecular systems are built up of classical, quasi-classical and quantum mechanical components.[3] This is because the limiting factor in computational quantum chemistry, like in many other computational sciences, is the time needed to process an underlying model. Owing to the high dimensionality of the respective Schrödinger equation, it would in practice simply be intractable to numerically calculate the time evolution of most molecular systems using only quantum mechanics.

Thus, the leading idea of virtually all molecular dynamics approaches is to treat the electronic problem by solving the Schrödinger equation whereas the motion of the nuclei is calculated through classical mechanics. The overall idea behind this strategy is that we begin, for instance, with relativistic quantum dynamics at the bottom level and then advance, firstly, to quantum dynamics via the approximation of small velocities, secondly, to atomic quantum dynamics via the approximation that electrons move much faster than nuclei, and, finally, to molecular dynamics via the approximation that the atomic motion is classical.

[3] For a more detailed discussion of the different models employed in quantum chemistry see Seck (2012).

The ascent to levels that employ classical motions of the atoms ultimately allows for using so-called particle-models. Such particle-models that describe systems in terms of particle interactions are the basis for many common computer simulations. These particles need not be 'small': proper particles are, for instance, nuclei, atoms, molecules, stars, or even clusters of galaxies. A computer simulation of particle models, i.e. running an algorithmic implementation of the underlying particle model, has to determine the resulting forces for each particle in each dynamical step.

Therefore, in the course of modelling molecules as interacting atoms which are composed of nuclei and electrons we must – at least in principle – solve the Schrödinger equation with the respective Hamiltonian. Because computational power (i.e. floating point operations per second and fast enough memory) is limited, simulationists always aim to optimize the algorithmic complexity by describing the atomic interactions in a way that offers an implementation that minimizes computational costs. As already indicated, virtually all molecular dynamics modelling treats the electronic problem by approximately solving the electronic Schrödinger equation to obtain the effective potential energy of the nuclei at each molecular dynamics step, whereas the motion of the nuclei is calculated using classical mechanics.

Within the so-called Ehrenfest molecular dynamics, for instance, Newton's law of force is coupled with the electronic Hamiltonian. The nuclei themselves are treated as classical particles. An alternative to the Ehrenfest molecular dynamics is the so-called Born-Oppenheimer molecular dynamics, where one additionally considers the large mass difference between the electrons and the nuclei, which entails that the velocities of the electrons be considerably higher than those of the nuclei. Furthermore, it can be assumed that the electrons are, at any time, in their ground state. In contrast to the Ehrenfest molecular dynamics, in the Born-Oppenheimer case, only a static, time-independent electronic structure problem needs to be solved in each dynamical step.

With respect to the implementation process, let me only suggest here that any molecular dynamics employs a mapping of mathematical symbols on algorithms and data structures and, finally, on binary code operating on CPU registers, which are allocated with values of physical properties (e.g. nuclear positions and velocities). The output, then, is a data structure filled in particular with the nuclear position and velocity of each nucleus at each (discrete) time step. To illustrate this point, I shall present an example.

If quantum chemists simulate, for instance, the rupture process of a molecule that is initially chemically anchored on a suitable substrate (say, a gold surface) in the course of an atomic force microscopy experiment (the molecule is subsequently pulled off the surface by an external force), they obtain a data structure tracing inter alia the nuclear positions at each time step between the bound initial and unbound final state of the molecular system. Finally, the output can be presented – using, for instance, the *GaussView 5* graphical interface of the Gaussian package – as one of the highly impressive (animated) representations of the time evolution of a system under investigation.

In the next section I shall consider the question whether such representations – and the generating equations respectively – can be counted as a basis for suitable scientific explanations.

3 Explanation in Quantum Chemistry

If we consider the highly idealized and hybrid (i.e. quantum-classical) character of the models of quantum chemistry, it seems simply impossible – at least strictly speaking – to interpret the representation given by the data structures generated through computer simulations as full-fledged scientific explanations of the behaviour of real-world systems. Michael Weisberg recently rightly characterizes the philosophical status of chemical modelling as follows (Weisberg 2012, p. 357):

> The point of many exercises in chemical modelling is to learn about real systems. In these cases, the model must bear certain relationships to real-world systems. Since chemical models are always incomplete, we cannot simply say that models are "true of" chemical systems.

Furthermore, he accordingly characterizes molecular orbital and valence bond models as highly idealized with respect to a full-fledged quantum mechanical treatment of a chemical system because in "their simple forms they neglect aspects of electron/electron interactions, relativistic effects, and correlation, all of which make considerable difference to the properties of some molecular systems" (Weisberg 2012, p. 357).

With respect particularly to the computational costs of computer simulations performed in quantum chemistry, Paul Humphreys recently arrived at the closely related conclusion that many idealizations and approximations are required because "*ab initio* calculations of the energy levels are impossible to carry out for any but the smallest molecules" (Humphreys 2009, p. 624). To broaden the scope of *ab initio* calculations quantum chemists typically employ the so-called Hartree-Fock self-consistent field approach utilizing approximations that are "inextricably linked with the degree to which those calculations can actually be carried out in practice" (Humphreys 2009, p. 624).

If reductionist minded chemists would like to deal with this situation they could defensively maintain that idealizations in general and the Ehrenfest approximation of the nuclei (i.e. the rationale for treating nuclei as classical particles) and the Hartree-Fock approximation (which employs, for instance, only the average repulsions between electrons) in particular will ultimately vanish within the framework of a de-idealizing research program. Philip Kitcher, for instance, has famously sketched such a program as early as 1989. After exploring the history of the concept of the chemical bond he argues in his well-known paper on explanatory unification in favour of de-idealization. While discussing reductive stages from pure chemical concepts like "valence" by way of physicochemical concepts like the "shell model of the atom" he ended up with full-fledged explanations of

bond-formation as a consequence of the stability of quantum mechanical systems. Then, he ultimately presumed that, although "this is only mathematically tractable in the simplest examples, it does reveal the ideal possibility of a further extension of our explanatory derivations" (Kitcher 1989, 447).

In order to evaluate such a claim, let us take a look at recent topical research in quantum chemistry. Although there are many methods and techniques like, for instance, semi-empirical methods, the Hartree-Fock approach arguably yields the only account that seems to be compatible with the aforementioned *Euclidean ideal*. Therefore, a further extension of our explanatory derivations (meaning a still further de-idealized quantum chemistry) would have to get rid of the Hartree-Fock approximation as well. At first sight, quantum chemists have in fact already begun working in this direction. In a recent textbook on *ab initio* molecular dynamics, we can yet find the following appraisal (Marx and Hutter 2009, p. 74 f.):

> Although post Hartree-Fock methods have a very unfavourable scaling of the computational cost as the number of electrons increases, a few case studies were performed with such correlated quantum chemistry techniques. [...] It should be kept in mind that the rapidly growing workload of post-HF [Hartree-Fock] calculations, although extremely powerful in principle, limits the number of explicitly treated electrons to only a few.

Generally, however, there is no hint that the models of quantum chemistry are so-called 'Galilean idealizations', a concept originally coined by McMullin (1985) which, in the long run of exhaustive de-idealizing aim "to give complete, non-distorted, perfectly accurate representations" (Weisberg 2007, 657) of a target system. Rather, post Hartree-Fock theories are driven by efforts to simulate systems to which the Hartree-Fock approximation simply cannot be applied. Furthermore, there seems to be a trend in the practice of quantum chemistry to employ a variety of different methods to explore one and the same system (Marx and Hutter 2009, p. 417):

> Furthermore, the mode of most applications has shifted principally in the new millennium away from feasibility and proof-of-principle studies at the cutting edge toward systematic and large-scale investigations. This means that properties and processes can now be studied comprehensively by varying several control parameters such as temperature, pressure, chemical composition, etc. simultaneously, as known for a long time from traditional computer simulations. At the same time these new possibilities allow one to check more carefully than ever before for statistical as well as systematic errors in the results, for instance by studying the same system using more than one electronic structure method by employing, for example, different density functionals.

These indications might be taken to suggest that there is no straightforward reduction of the plurality of models, methods and techniques via de-idealizing to be found even in the practice of current quantum chemistry. Quite on the contrary, conducting multiple calculations aiming at the same target system using different approaches and models seems to be a likewise fruitful and common heuristic in current research to produce and validate explanations.

4 Conclusion

It seems to me a bad option for philosophers of science to take allegedly de-idealizing research strands that are apparently not distinctly evidenced in practice as an inducement to characterize and justify chemical explanations. Alternatively, we shall take some further explanatory virtues seriously. If quantum chemists try to give, for instance, an explanation for the time-evolution of a cluster of molecules, the most important question for them is: which model and which approximation method is the most efficient in terms of computational cost given a specific target system? Additionally, quantum chemists consider first, the range of the involved parameters (energies, forces, etc.) and, second, the quantity of systems to which the model is applicable. Thus, a tentative conclusion would be that the hybrid models of quantum chemistry should not be counted as genuinely explanatory only (if at all) because of their (more or less airy) embedment within quantum mechanics. By contrast, the explanatory status of these models may better be characterized and justified by a certain set of properties. This set may include, inter alia, the range of applicability to different target systems, the algorithmic complexity and the models partaking in an autonomous chemical research framework.

References

Dirac, P. (1929). Quantum mechanics of many-electron systems. *Proceedings of the Royal Society London, 123*, 714–733.

Gaussian. (2014). Gaussian 09: Expanding the limits of computational chemistry. http://www.gaussian.com/g_prod/g09b.htm

Gavroglu, K. (1997). Philosophical issues in the history of chemistry. *Synthese, 111*, 283–304.

Gavroglu, K., & Simoes, A. (2012). *Neither physics nor chemistry. A history of quantum chemistry*. Cambridge: MIT Press.

Hendry, R. F. (2012). Reduction, emergence and physicalism. In A. Woody & R. Hendry (Eds.), *Philosophy of chemistry* (pp. 367–386). Amsterdam: Elsevier.

Hendry, R. F., & Needham, P. (2007). Le poidevin on the reduction of chemistry. *British Journal for the Philosophy of Science, 58*, 339–353.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*, 615–626.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 410–504). Minneapolis: University of Minnesota Press.

Marx, D., & Hutter, J. (2009). *Ab initio molecular dynamics. Basic theory and advanced methods*. New York: Cambridge University Press.

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

Needham, P. (2010). Nagel's analysis of reduction: Comments in defence as well as critique. *Studies in History and Philosophy of Modern Physics, 41*, 163–170.

Scerri, E. (1997). The periodic table and the electron. *American Scientist, 85*, 546–553.

Scerri, E. (2004). Just how ab initio is ab initio quantum chemistry. *Foundations of Chemistry, 6*, 93–116.

Scerri, E., & McIntyre, L. (1997). The case for the philosophy of chemistry. *Synthese, 111*, 213–232.

Seck, C. (2012). Metaphysics within chemical physics: The case of ab initio molecular dynamics. *Journal for General Philosophy of Science, 43*, 361–375.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy, 104*, 639–659.

Weisberg, M. (2012). Chemical modeling. In A. Woody & R. Hendry (Eds.), *Philosophy of chemistry* (pp. 355–363). Amsterdam: Elsevier.

Are Chemical Kinds Natural Kinds?

Robin Findlay Hendry

1 Introduction

Philosophers of science have long explored the role of natural kinds in the core business of science: classifying the world, and understanding the processes that underlie the possibility of successfully predicting and manipulating it (Hacking 1991, 2007). More recently, kinds have been part of a philosophical project that has come to be allied explicitly with scientific realism. Chemical kinds were central examples: Saul Kripke (1980) and Hilary Putnam (1975) chose gold and water as paradigms respectively, arguing that semantic access to these natural kinds could be independent of detailed theoretical knowledge of them, in a way that made them (for Putnam) the neutral subjects of successive theories in the history of science, even those making mutually incompatible claims about the very same kinds. The aim was to show how it is possible for natural kinds to be discovered rather than imposed on nature. It is difficult to think of a single aspect of either Kripke's or Putnam's projects, from semantic externalism to the grounding of metaphysical claims in semantic facts, that is not highly controversial. Nevertheless, according to Brian Ellis their work did have the effect of making 'belief in essences or essential natures once more respectable' (2002, 7), at least for those philosophers who were willing to pursue speculative projects in metaphysics. In this spirit, Ellis constructed a unified account of the metaphysics of science prominent in which is a highly restrictive view of natural kinds. Like Kripke and Putnam, Ellis took chemical kinds as his model arguing that, unlike biological species, they are hierarchically ordered, discrete, and independent of human interests. Thus in Ellis' account, only some categories that underlie reliable generalisations are natural kinds. In contrast, a more

R.F. Hendry (✉)
Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK
e-mail: r.f.hendry@durham.ac.uk

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_19

liberal (but still realist) conception of kinds would remake the link with scientific generalisation, seeking only the objective properties that underlie it, without any *a priori* assumptions about the relationships between them, or the kinds that they generate. Although I am sympathetic to this more liberal view, the main point of this paper is not to argue positively for it, but rather to establish that not even chemical kinds fit the more restrictive view that claims to take them as exemplars. The more restrictive view is untenable if chemical substances are to occupy their traditional place as central exemplars of the notion of a natural kind.

2 Identifying the Question

To answer the central question of this paper, I first need to identify what that question is. What are chemical kinds, and what, in general terms, is it to be a natural kind?

2.1 Chemical Kinds

Chemical kinds are just the general categories that appear in chemical theories and explanations. It is easiest to identify them by example. Chemistry is concerned with the properties and behaviour of chemical substances, and explaining both in terms of their structure at the molecular scale. So chemical kinds include both substances (and various higher kinds of them) and such microstructural items as atoms, molecules and ions (and various higher kinds of *them*). Substances may be elements, compounds or mixtures. The elements, of which there are something over a hundred, are those substances that have no others as chemical components. Chemists arrange them into groups based on their patterns of chemical reactivity (such as the alkali metals and halogens), and also their material structure in the free state (such as the metals). Compound chemical substances are named and classified by their structures (Hendry 2006, 2008, forthcoming), and also grouped together in classes that share either an elemental component (e.g. chlorides), a microstructural feature (e.g. carboxylic acids), or merely a pattern of chemical reactivity (e.g. acids).

2.2 Kinds of Naturalness

In that part of the philosophy of science which is closest to metaphysics, discussions of natural kinds are concerned with the naturalness of *groupings* of objects or substances. Historians and philosophers who are more closely concerned with scientific practice—and less engaged with metaphysical issues—sometimes point out that chemistry is a synthetic science whose practical achievements consist, to a large extent, in the making of new (artificial, *non*-natural) kinds of stuff. There is a

real danger of confusion here. Chemical substances themselves may or may not be naturally occurring, but that has nothing to do with the naturalness of the category they fall under. Take for example fluorine: its status as an element was inferred long before it was isolated. Lavoisier listed 'fluoric radical' in his table of simple substances (Lavoisier 1790, 175), and the name 'fluorine' was introduced by André-Marie Ampère in 1812 (Greenwood and Earnshaw 1984, 920). Mendeleev afforded fluorine a place in the earliest versions of the periodic table in the 1860s, grouping it with the other halogens chlorine, bromine and iodine (Mendeleev 1869/2002). Yet a reliable method for isolating fluorine safely was developed by Henri Moissan only in 1886. There was such a long gap between the inference and the isolation simply because fluorine gas does *not* occur naturally: it reacts explosively with glass and water, and virtually anything else. But in the minds of chemists (and also in nature) fluorine gas seems to have been a natural category awaiting developments in human artifice to populate it. The distinction between natural and artificial substances is not without interest, but the most important things to say about it are surely sceptical: that it is a parochial distinction, and of no global significance. Why? Because it must depend on the particular chemical environments and processes that govern the production and survival of substances in the local region surrounding the person who utters the word 'natural'. From a global point of view, any substance that the laws of nature allow to be made is, in a sense, natural.

It is a straightforward (category) mistake to confuse the two kinds of naturalness, but philosophers, historians and sociologists who are aware of the distinction sometimes do not keep them keep them far enough apart. Thus, for instance, Nalini Bhushan (2006, 328) correctly distinguishes the issue of 'natural versus arbitrary groups of objects' from the question of whether something is '"naturally occurring" (found in nature)', but she then argues that chemistry's involvement in synthesis, sometimes of exotic new substances, undermines (simple versions of) the idea that chemical kinds can be said to have been discovered. But there are two kinds of discovery here: in physics, the discovery of a natural phenomenon may involve the invention of an apparatus that will display it (Buchwald 1994 provides a beautiful example in the case of Heinrich Hertz and electric waves). What Moissan and Hertz discovered was how to produce something that falls under a natural category (fluorine, or electric waves). The categories had already been discovered.

Philosophers sometimes distinguish natural kinds from artifactual kinds, giving tables and chairs as examples of the latter. But there is a danger of confusion here too. What makes chairs and tables 'non-natural' is not just the fact that they were made by human hands. 'Table' and 'chair' also name functional kinds that bear no simple relationship to any natural properties. Substances can be described in functional terms too: many different chemical substances can act as analgesics, for instance. Philosophers also sometimes distinguish natural kinds from random collections: 'Natural kinds are standardly distinguished from arbitrary groups of objects, such as what you had for breakfast' (Daly 2011). This characterisation is criticised by Ian Hacking, because when eighteenth-century naturalists worried about the naturalness of the Linnaean system, '[c]lasses were artificial rather than natural, when they had been invented by botanists, but did not accurately represent

the order of living things' (Hacking 2007, 211). Some conventionalists elide the difference between natural and other categories by showing that even those which are natural to us are interest-relative. All such categories, they argue, are invented or imposed on nature. Others argue that nature does not to favour this category over that, so the choice is arbitrary. So either 'arbitrary' or 'artificial' can be used to make a contrast with naturalness, and to deny the existence of one. Given our earlier discussion, it is surely best to avoid the word 'artificial' when discussing synthetic sciences. Either way, being a realist about natural kinds involves taking on the task of distinguishing natural from non-natural categories in some way that grounds the distinction in the order of nature. I haven't yet given a positive account of how to do that. I will do so in Sect. 4, but for the moment I would like to critically examine the restrictive view.

3 Chemical Kinds as Natural Kinds

Do the chemical categories, those paradigms of natural kinds, satisfy the three necessary conditions that characterise a restrictive view of kinds?

3.1 Hierarchy

The hierarchy requirement is that no two natural kinds may overlap (have instances in common) unless (on the strong version) one includes the other, or (on the weak version) both are included within a third (see Hacking 1991, 2007; Khalidi 1998, 2013; Ellis 2001, 2002, 2009; Tobin 2010). The broad motivation is that natural kinds should form a single unified system: a nested hierarchy. But why should they? Unity, of which hierarchy is one aspect, might well be a virtue of a system of classification. Here the motivation is presumably different for realists and anti-realists about classification. Anti-realists hold that classification is just science imposing on the world categories that fit its pragmatic or theoretical interests. If unity is just one more interest, the only question is how it trades off against others. For this very reason, it seems wrong for it to be a hard-and-fast requirement. For realists I am even less sure about the motivation. It might be thought that the requirement for unity in a system of kinds reflects the broad realist view that science investigates one world. But one world could contain many systems of kinds, and why must a unified system of kinds be a hierarchical one? One might instead see unity as arising from a single *basis* for classification, such as microstructure in chemistry. Such a basis may or may not give rise to a nested hierarchy of kinds.

A brief look at chemistry undermines the stronger version of the hierarchy requirement (the weaker one might be very weak, as we shall see). Emma Tobin points out (2010, 189) that there is an overlap between tin and the metals, but neither encompasses the other. Tin comes in two common forms, or allotropes:

white (metallic), and grey (non-metallic). The transformation from white to grey begins to happen spontaneously as the temperature falls below about 13 °C. Objects made of white tin, such as buttons or fuel containers, disintegrate in the process and are said to have succumbed to 'tin pest' or 'tin blight'. Defenders of the hierarchy requirement might challenge either of these categories as natural kinds, but neither challenge seems convincing. Tin is an element, and what makes something tin is its having a particular nuclear charge (50 atomic units), a property which explains tin's particular profile of chemical and physical behaviour. What makes something a metal is its having a particular structure: an array of metal ions surfed by electrons which are relatively free because they are only loosely bound to the atoms. This structure is what explains the common properties of metals, such as their electrical and thermal conductivity, and their lustrous appearance. It is a *prima facie* objection to any approach to natural kinds that it forces us to reject either of these categories. One might instead move to a weaker hierarchy requirement (in fact there are many: see Tobin 2010 for the distinctions), appealing to the elements as a natural kind that encompasses both tin and the metals. This appeal fails because the elements do *not* encompass both tin and the metals: some metals (alloys such as steel and brass) are not elements. What about 'chemical substance' as the encompassing category? I have serious doubts about 'chemical substance' as a natural kind, partly because nature does not seem to provide a foundation for the distinction between pure substances and mixtures, in either macroscopic or microscopic terms. Without any such foundation (and, more importantly, a contrast), appeal to the category of substances is empty: a cheap way of saving the hierarchy requirement. Far better to drop a requirement whose motivation is anyway unclear than defend it in this disreputable way.

3.2 Discreteness

The idea that kinds must be discrete has a long history: Plato used the metaphor of 'cutting nature at its joints' to argue that some ways of categorising the world are more natural than others. Locke worried that where individuals of one kind can be transformed continuously into individuals of another, some might be on the cusp between them: 'monsters' that would defy the distinction (see Lowe 2011, 5–6). Ellis shares the worry, and assumes that kinds must be discrete:

> Natural kinds exist if and only if there are objective mind-independent kinds of things in nature. Hence, to believe in natural kinds one must believe that things are divided naturally into categorically distinct classes. (2009, 57)

Chemical substances and animal species, he thinks, provide two contrasting cases:

> The elements and their various compounds are all *categorically distinct* from each other. They are distinct in the sense that there is never a gradual transition from any one chemical kind to any other chemical kind. Consequently, it is never an irresolvable issue to which chemical kind a given chemical substance belongs. Where there are such transitions in nature, as there are between the colours, for example, we have to draw a line somewhere if we wish to make a distinction. (2002, 26)

Elsewhere he argues that 'There is no continuous spectrum of chemical variety that we had somehow to categorize,' although '[w]hat is true of the chemical kinds is not true of biological species. The existing species of animals and plants are clusters of morphologically similar organisms' (2009, 59). Against both Locke and Ellis, I will argue that neither the mere possibility nor the actual occurrence of continuous transformations between two chemical kinds entails that they fail to be objectively distinct. With the exception of the elements, chemical variety is continuous.

Firstly, continuous transition between distinct chemical substances is possible at least in thought. I have argued elsewhere that chemistry individuates substances by their microstructures (Hendry 2006, 2008; see Needham 2000, 2011 for the contrary view). The microstructures of compound substances themselves seem to be defined by continuously varying quantities such as distances between atoms.[1] Secondly, such continuous transition between distinct chemical species is exactly how theoretical explanations depict chemical transformation. William Goodwin (2012) distinguishes two notions of reaction mechanism: in what he calls the 'thick' sense, a mechanism traces the motions of the constituent atomic nuclei and electrons in a continuous path from the reagents' molecular structure to that of the products; a mechanism in the 'thin' sense neglects some parts of the process in order to concentrate on a few theoretically important and well-understood stages. How the two conceptions fit together is a good question (Goodwin 2012, 310–15), but in either case the transitions are continuous. In thick mechanisms this is explicitly the case, but it is also true of the steps that make up a thin mechanism. Consider for instance a textbook example: the reaction of an alkyl halide RX (for instance bromoethane) with a nucleophilic ion Nuc^- (for instance the hydroxyl ion OH^-):

$$RX + Nuc^- \rightarrow RNuc + X^-$$

Depending on the nature of the alkyl group R, the 'leaving group' X^-, and the conditions under which the reaction takes place (e.g. the nature of the solvent), the reaction proceeds via one of two mechanisms, called S_N1 and S_N2 (standing for unimolecular and bimolecular nucleophilic substitution respectively). For kinetic purposes the S_N1 reaction is regarded as taking place via two steps: a (slow) unimolecular dissociation of the alkyl halide:

$$RX \rightarrow R^+ + X^-$$

followed by a rapid attack by the nucleophile:

$$R^+ + Nuc^- \rightarrow RNuc$$

[1] In fact the notion of a structure is a little more complicated, because one must distinguish geometrical structure from bond structure: see Hendry 2013, and below.

The slow first step involves the breaking of a bond between the halogen atom and the neighbouring carbon atom in the alkyl group. Isn't the breaking of a bond a discontinuous process? Not if it involves a gradual lengthening and weakening. It all depends on what a bond is. On one influential account (Bader 1990), bonds are topological features (defined in terms of local maxima and minima) of a molecule's electron-density distribution. If so, then the making and breaking of bonds consists in continuous rearrangements of electron density. The apparently discontinuous bond-breaking arises from underlying continuity, much as the peak in a mathematical function can be made to disappear discontinuously by continuous transformation of the function of which it is a peak.

The third argument draws on the idea of a potential energy surface (PES), which is a contour map in which each point is a distinct geometrical configuration of a molecule, and the height of the surface at that point is the energy of that particular configuration. Stable (i.e. low-energy) configurations correspond to minima on the PES (one might think of a valley). Transitions between stable substances typically require the scaling of an energy barrier: a journey over a mountain, or through a mountain pass, to a neighbouring valley. Once again, chemical change is presented as gradual transition from one chemical kind to another, so chemical kinds cannot be regarded as discrete. Interestingly, in transforming from one stable configuration to another, a particular molecular system may fleetingly inhabit a high-energy transition state (a Lockean 'monster'), but this presents no classificatory antinomies. Chemists seem to be happy that the regions of a structural configuration space corresponding to different chemical substances may overlap and have vague boundaries. They are surely right that these facts have no tendency to undermine the objective distinctness of those regions, or of the corresponding chemical substances.

As with the hierarchy requirement, one might quibble about the status of natural kinds that violate the discreteness condition. But such a defence of the discreteness requirement is surely disastrous: apparently, no compound substances are to count as natural kinds. I take it that we should turn away from any account of natural kinds which has this consequence. Different chemical compounds are readily separable by natural means. The differences between them are recognised by nature's laws, and so are part of nature itself.

3.3 Independence from Interests

That natural kinds should be independent of the interests of classifiers is an equally venerable requirement. Once again it is applied both in both defence and criticism of the idea that chemical kinds are natural kinds. Thus Ellis 2002 (26–27) argues that 'questions of chemical identity can never be dependent on our interests, perceptual apparatus, psychologies, languages, practices or choices' (2002, 26). In contrast Donellan (1983, 98–104) worries that chemists' interest in atomic number over atomic weight results from their 'psychological quirks' (1983, 103) and 'historical accident' (1983, 104). Achille Varzi argues that 'Even in physics, our microscopic

categories seem to suffer from a variety of human contingencies' (2011, 141), although his example is the same as Donellan's: isotopy and the chemical elements. While the *existence* of classes of atoms of like nuclear charge (or of like atomic weight) is clearly not interest-dependent, chemists' focus on nuclear charge rather than atomic weight clearly is. I agree with Donellan and Varzi that chemical kinds are interest-dependent in important ways, but do not think that this in any way undermines their status as natural kinds. Elsewhere (2006, 2008, 2010), I have argued that historical scientists such as Lavoisier should be understood as using the names of elements as referring to classes of atoms or ions of like nuclear charge even though they were agnostic about the existence of atoms and were entirely innocent of the notion of nuclear charge, because nuclear charge is the main determinant of the kinds of chemical behaviour they were seeking to understand. Atomic weight is a much less important factor, and many substances like mercury, silver and tin that were known as elements in Lavoisier's time are isotopically diverse (that is, heterogeneous in respect of their atomic weight), as were the samples they had of these elements. What of compound substances? Chemistry individuates them by their microstructures. While I think that it would be incorrect to regard this as a *choice* that the discipline has made, alternatives are available. Other disciplines and activities pick out substances from the point of view of quite different interests: wool and silk are individuated by the biological species in which they originate, and it is well known that 'jade' refers to two distinct chemical substances (see LaPorte 2004). These are all feasible ways of individuating substances, but they are not *chemistry's* ways. Structure infects the discipline's entire approach to substances and the important similarities and differences between them. But the interest-dependence does not end there (see Hendry 2013): 'structure' itself means a number of different things, depending on context. First distinguish the bond structure of a substance (the order of bonded connections between its atoms) from its geometrical structure (the relative positions of atoms and ions in a substance). Neither, I argue, is more fundamental than the other, and geometrical structure itself is different at different (energy, length or time) scales. Any substance has different structures, and which one is relevant depends on what one is trying to explain. In short, structure, and therefore chemical classification, is interest-dependent.

4 Liberal Kinds

A more liberal approach to the notion of a natural kind starts with the practice of making generalisations and predictions: natural kinds are based on genuine similarities that underwrite scientific understanding, and the non-trivial explanation of systematically successful prediction in science. This is central to the broad tradition of natural kinds in the philosophy of science, as Hacking traces it (2007). On this broad conception, a kind may be natural in so far as its members share some property[2] which is causally relevant to maintaining or explaining the regularities in

[2]Or, to accommodate homeostatic property clusters, the same vague-boundaried region of a property space.

behaviour whose existence underlies the very usefulness of the notion of a natural kind. This approach, with which I have much sympathy, ought to be inclusive, and hence pluralistic like John Dupré's promiscuous realism (1993), or Muhammad Ali Khalidi's naturalistic view of kinds and classification (2013): there are many natural kinds, corresponding to all the different ways in which objective differences are causally relevant to understanding the similarities and differences in the behaviours of things. Moreover there are no particular reasons to think that natural kinds should bear any particular relationships to each other, or form a single system. It is possible that they do, but that is a matter of how the world is, and therefore for empirical research. The formation of classificatory *systems* is undeniably an important part of science (see Hendry 2012), but arises as a feature of highly developed scientific disciplines. It cannot be imposed *a priori*. Human kinds, artifactual kinds and even functional kinds can all be natural, on this view, although there may be differences among them depending on the depth (or non-triviality) of the scientific explanations, and the explanation of the systematically successful predictions they are involved in. None of this implies that different kinds, or classificatory systems, are on a par: there are limits to pluralism (Hendry 2012).

This approach gives us a quick answer to the question of whether or not chemical kinds are natural: they are, because they are good bases for successful explanation and prediction in a particular empirical science—chemistry—and it is quite clear how they are able to do this. In some cases, as in the elements, I believe a full-blown essentialist explanation is available (sketched in Hendry forthcoming). In others, such as the acids, the explanations are much thinner: there is probably nothing more to being an acid than displaying enough of the typically acidic kinds of behaviour. The explanation is thinner because although there is a causal explanation of why any particular acid behaves the way it does, the explanations for different acids diverge. No single component or structural property gives rise to all the different cases of acidity.

5 Conclusion

Are there any conditions that a scientific category *must* satisfy, if it is to qualify as a natural kind? Given the foregoing, one might answer no, and go on to reject the very notion of a natural kind, because it can have no useful role to play in understanding how science works. Ian Hacking (2007) does just this, worrying that it has been debased by the sheer variety of conflicting uses to which philosophers have put it. Moreover, he rejects a stipulative revision that 'picks out some precise or fuzzy class and defines it as the class of natural kinds' (2007, 238). I tend to agree on the diagnosis, but reject the prognosis: just because it fails to admit of necessary and sufficient criteria, or a clear *region* of applicability (even with vague boundaries), it does not follow that the term 'natural kind' is useless. It is no better or worse off in this respect than many natural-kind terms which nevertheless serve a useful purpose in science without stipulative redefinition to settle discussion in favour of one competing conception.

Interestingly, 'acid' is one such example. There are core examples in the strong acids: nitric acid, hydrochloric acid and sulphuric acid. The behaviour of these three cases is genuinely similar in important respects, and so are the mechanisms by which the acidic behaviour arises. The problem is how extend the category from these core examples to provide necessary and sufficient criteria for acidity, whether in compositional, structural or functional terms. In the history of chemistry, all such attempts failed (see Finston and Rychtman 1982, Chapter 1). Kyle Stanford and Philip Kitcher (2000) use this example to argue for a revision of Putnam's account of natural kinds, but it is hard to see why the behaviour of one kind term should undermine the much stronger cases for its applicability to other cases such as, for instance, the elements, that might be made on the basis of chemistry and its history (see Hendry 2010, forthcoming). From this multi-level story I take two lessons.

Firstly, arguments for or against particular semantic or metaphysical claims about natural kinds ought to be made out on a case-by-case basis. The idea of taking a *general* essentialist position about natural kinds, or indeed a *general* externalist position about natural kind-terms, is silly, unless it means being an essentialist or an externalist about at least one natural-kind term. Secondly, a name for a general phenomenon, kind of behaviour or syndrome can be useful even if, it turns out, there is no unified explanation for that general phenomenon, kind of behaviour or syndrome. *Pace* Hacking, one can think that natural kinds are a serious topic for philosophical discussion without thinking that there are necessary and sufficient criteria for being a natural kind. One can even be an essentialist about some natural kinds without being an essentialist about 'natural kind' (cf. Dupré 2002).

Moreover, from all the foregoing I have drawn (what I take to be) an informative conclusion about natural kinds: that they need not be hierarchically related to other natural kinds, discrete, or interest-independent.

Acknowledgements I would like to thank Tuomas Tahko for organising an excellent symposium on 'Debating Natural Kinds' at the EPSA meeting in Helsinki in August 2013. I would also like to thank Jordan Bartol and two anonymous referees for comments and suggestions.

References

Bader, R. (1990). *Atoms in molecules: A quantum theory*. Oxford: Oxford University Press.

Bhushan, N. (2006). Are chemical kinds natural kinds? In D. Baird, E. Scerri, & M. I. Lee (Eds.), *Philosophy of chemistry: Synthesis of a new discipline* (pp. 327–336). Dordrecht: Springer.

Buchwald, J. (1994). *The creation of scientific effects: Heinrich Hertz and electric waves*. Chicago: The University of Chicago Press.

Daly, C. (2011). Natural kinds. In E. Craig (Ed.), *Routledge encyclopedia of philosophy*. London: Routledge.

Donnellan, K. (1983). Kripke and Putnam on natural kind terms. In C. Ginet & S. Shoemaker (Eds.), *Knowledge and mind: Philosophical essays* (pp. 84–104). Oxford: Oxford University Press.

Dupré, J. (1993). *The disorder of things*. Cambridge: Harvard University Press.

Dupré, J. (2002). Is 'natural kind' a natural kind term? In *Humans and other animals* (pp. 103–123). Oxford: Oxford University Press.
Ellis, B. (2001). *Scientific essentialism*. Cambridge: Cambridge University Press.
Ellis, B. (2002). *The philosophy of nature: A guide to the new essentialism*. Chesham: Acumen.
Ellis, B. (2009). *The metaphysics of scientific realism*. Durham: Acumen.
Finston, H. L., & Rychtman, A. C. (1982). *A new view of current acid–base theories*. New York: Wiley.
Goodwin, W. (2012). Mechanisms and chemical reaction. In R. F. Hendry, P. Needham, & A. I. Woody (Eds.), *Philosophy of chemistry* (pp. 309–327). Amsterdam: Elsevier.
Greenwood, N. N., & Earnshaw, A. (1984). *Chemistry of the elements*. Oxford: Pergamon Press.
Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies, 61*, 109–126.
Hacking, I. (2007). Natural kinds: Rosy dawn, scholastic twilight. *Royal Institute of Philosophy Supplement, 61*, 203–239.
Hendry, R. F. (2006). Elements, compounds and other chemical kinds. *Philosophy of Science, 73*, 864–875.
Hendry, R. F. (2008). Microstructuralism: Problems and prospects. In K. Ruthenberg & J. van Brakel (Eds.), *Stuff: The nature of chemical substances* (pp. 107–120). Würzburg: Königshausen und von Neumann.
Hendry, R. F. (2010). The elements and conceptual change. In H. Beebee & N. Sabbarton-Leary (Eds.), *The semantics and metaphysics of natural kinds* (pp. 137–158). London: Routledge.
Hendry, R. F. (2012). Chemical substances and the limits of pluralism. *Foundations of Chemistry, 14*, 55–68.
Hendry, R. F. (2013). The metaphysics of molecular structure. In V. Karakostas & D. Dieks (Eds.), *EPSA11 perspectives and foundational problems* (pp. 331–342). Berlin: Springer.
Hendry, RF. (forthcoming). Natural kinds in chemistry. In G. Fisher & E. Scerri (Eds.), *Handbook of the philosophy of chemistry*. Oxford: Oxford University Press, forthcoming.
Khalidi, M. A. (1998). Natural kinds and crosscutting categories. *Journal of Philosophy, 95*, 33–50.
Khalidi, M. A. (2013). *Natural categories and human kinds*. Cambridge: Cambridge University Press.
Kripke, S. (1980). *Naming and necessity*. Cambridge: Harvard University Press.
LaPorte, J. (2004). *Natural kinds and conceptual change*. Cambridge: Cambridge University Press.
Lavoisier, A. (1790). *The elements of chemistry* (trans: Robert Kerr). Edinburgh: William Creech. French edition: Traité Elémentaire de Chimie. Paris: Vrin.
Lowe, E. J. (2011). Locke on real essence and water as a natural kind: A qualified defence. *Aristotelian Society Supplementary, 85*, 1–19.
Mendeleev, D. I. (1869/2002). On the relation of the properties to the atomic weights of the elements. In W. B. Jensen (Ed.), *Mendeleev on the periodic law: Selected writings, 1869–1905* (pp. 16–17). New York: Dover.
Needham, P. (2000). What is water? *Analysis, 60*, 13–21.
Needham, P. (2011). Microessentialism: What is the argument? *Noûs, 45*, 1–21.
Putnam, H. (1975). The meaning of 'meaning'. In *Mind language and reality* (pp. 215–271). Cambridge: Cambridge University Press.
Stanford, P. K., & Kitcher, P. (2000). Refining the causal theory of reference for natural kind terms. *Philosophical Studies, 97*, 99–129.
Tobin, E. (2010). Crosscutting natural kinds and the hierarchy thesis. In H. Beebee & N. Sabbarton-Leary (Eds.), *The semantics and metaphysics of natural kinds* (pp. 179–191). London: Routledge.
Varzi, A. C. (2011). Boundaries, conventions and realism. In J. K. Campbell, M. O'Rourke, & M. H. Slater (Eds.), *Carving nature at its joints: Natural kinds in metaphysics and science* (pp. 129–153). Bradford: MIT Press.

Part VI
Induction, Probability and Chaos

Why Bertrand's Paradox Is Not Paradoxical but Is Felt So

Zalán Gyenis and Miklós Rédei

1 The Main Claim

The paradox named after Bertrand was published first in Bertrand's book (1888). Bertrand asked the question: what is the probability that choosing a chord randomly in a circle, the length of the chord will be greater than the side of the equilateral triangle inscribed in the circle. The paradox is supposed to consist in the fact that parametrizing the random events involved in the situation in three different ways, one obtains three different probabilities of the event in question if one applies the Principle of Indifference to the different parametrizations. The standard interpretation of the paradox is then that it undermines the Principle of Indifference by showing it inconsistent and, to the extent the Classical Interpretation of probability is based on the Principle of Indifference, also the classical interpretation of probability. Bertrand's Paradox features in every book on probability theory that addresses interpretational issues and it continues to attract interest (recent

Research supported in part by the Hungarian Scientific Research Found (OTKA). Contract number: K83726.

Research supported in part by the Hungarian Scientific Research Found (OTKA). Contract number: K100715. The paper was completed while M. Rédei was an Honorary Research Fellow in the Institute of Philosophy of the Hungarian Academy of Sciences.

Z. Gyenis (✉)
Alfréd Rényi Institute of Mathematics, 13–15 Reáltanoda str. 1053 Budapest, Hungary
e-mail: gyz@renyi.hu

M. Rédei
Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK
e-mail: m.redei@lse.ac.uk

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_20

publications include Marinoff (1994), Mikkelson (2004), Shackel (2007), Bangu (2010), Rowbottom and Schackel (2010), Rowbottom (2013), and Aerts and de Bianchi (2014)).

The present paper suggests an interpretation of Bertrand's Paradox that is radically different from the standard one stated above. The details of our interpretation and the arguments supporting it have been published elsewhere (Gyenis and Rédei 2014), we only summarize here the main points. We claim in this paper that Bertrand's Paradox is not a paradox at all; quite on the contrary: properly formulated, the "paradox" simply states a provable, non-trivial mathematical fact, which is in perfect harmony with the correct intuition about how probability theory should be used to model random phenomena. Consequently, no "solution" of the paradox is needed and therefore all alleged solutions are misguided if they assume that there is indeed a paradox to be resolved in Bertrand's paradox. For this reason we do not aim at criticizing any of the alleged resolutions in detail. What is needed, in our view, is to explain why one may think and feel that there is something paradoxical in Bertrand's Paradox. Once this becomes clear, the paradox evaporates and so does the need to resolve it.

The basis of the interpretation suggested here is that the category of probability measure spaces with an *infinite* set of random events for which a classical interpretation of probability based on the **Principle of Indifference** can be meaningfully formulated is the one in which the set X of elementary events is a compact topological group and the probability measure p_H is the (normalized) Haar measure determined by the group on the Borel algebra $\mathcal{S}$ of subsets of X. To state the mathematical fact in question we define first a notion called **Labeling Invariance** in this category of measure spaces: **Labeling Invariance** states that a re-labeling $(X', \mathcal{S}')$ of random events $(X, \mathcal{S})$ is an isomorphism between $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ as probability measure spaces. **Labeling Invariance** can be interpreted as the expression (in the context of the classical interpretation) of the general intuition we call **Labeling Irrelevance**: that the specific way the random events are named is irrelevant from the perspective of the value of their probability.

Bertrand's Paradox is interpreted in this paper as the proposition that **Labeling Invariance** does *not* hold in the category of Haar probability measure spaces with an X of *infinite* cardinality. We will also argue however that **Labeling Invariance** is *not* the proper way to express **Labeling Irrelevance**: our freedom to choose measure theoretically isomorphic probability theories to describe the same random phenomenon manifests the conventionality of naming random events in probabilistic modeling; thus violation of **Labeling Invariance** is perfectly compatible with **Labeling Irrelevance**. It is the mistaken conflation of the correct intuition expressed by **Labeling Irrelevance** with the not maintainable **Labeling Invariance** which is the reason why one feels that Bertrand's Paradox is paradoxical.

2 The Elementary Classical Interpretation of Probability

The elementary classical interpretation of probability concerns the probability space $(X_n, \mathcal{P}(X_n), p_u)$, where X_n is a finite set containing $n < \infty$ number of elementary random events and the full power set $\mathcal{P}(X_n)$ of X_n represents the set of all events. The probability measure p_u is determined by the requirement that the probability $p_u(A)$ be equal to the "ratio of the number of favorable cases to the number of all cases", which is equivalent to saying that p_u is the probability measure that is uniform on the set of elementary events, which also is equivalent to saying that p_u is the probability measure that satisfies permutation invariance:

$$\text{for every } \pi \in \Pi_n \text{ one has: } p_u(\{x_i\}) = p_u(\{x_{\pi(i)}\}) \quad \text{for all } i \in \{1, 2, \ldots n\} \qquad (1)$$

where Π_n is the group of permutations of the n element set $\{1, 2, \ldots n\}$ and $\pi \in \Pi_n$ is a permutation.

It also is part of the classical interpretation what can be called the **Interpretive Link**: that the numbers $p_u(A)$ are related to something non-mathematical. Without such an interpretive link, the classical interpretation cannot be regarded as an *interpretation* of probability at all because the numbers $p_u(A)$ defined by (Eq. 1) are just mathematical stipulations. There are two standard **Interpretive Links**: The **Frequency Link** and the **Degree of Belief Link**. We restrict the discussion to the **Frequency Link** because the classical interpretation emerged historically and was formulated on the basis of this link. We will comment briefly on the **Degree of Belief Link** in the final section however.

Elementary Classical Interpretation: When the number of elementary events is finite, the probabilities of events are given by the measure p_u that is uniform on the set of elementary events, and (**Frequency Link**:) the numbers $p_u(A)$ will be (approximately) equal to the relative frequency of A occurring in a series of trials producing elementary random events from X_n.

Note that the reference to future events in the above formulation of the elementary classical interpretation distinguishes the classical interpretation (with the **Frequency Link**) from the frequency interpretation of probability, in which the ensemble of elementary random events determining A's relative frequency must be specified and fixed *before* one can talk about probabilities (cf. von Mises (1928, p. 24)).

It is clear that the classical interpretation formulated above is not maintainable however without some further restrictions: simple examples (such as flipping a biased coin) show that it is only under special circumstances that the number $p_u(A)$ is indicative of the frequencies with which A will occur in trials. This is what the **Principle of Indifference** is supposed to take care of:

Elementary Principle of Indifference: *If* the permutation group Π_n expresses epistemic indifference about the elementary random events in X_n, *then* the (Elementary) **Classical Interpretation** is correct.

Thus the (Elementary) **Principle of Indifference** states that the (elementary) classical interpretation of probability is maintainable if one is epistemically neutral in some sense about the elementary events; in Sect. 6 we will return to the issue of what the content of epistemic neutrality can possibly be.

3 The General Classical Interpretation of Probability in Terms of Haar Measures

Bertrand's Paradox type arguments, of which the original Bertrand's Paradox is but one example (Hájek (2012) formulates a few others), allegedly show that applying the **Principle of Indifference** is inconsistent because it can lead to assigning different probabilities to the same random event. Even a superficial look at the Bertrand's Paradox type arguments reveals that the paradoxes involve an (uncountably) *infinite* number of elementary events. But if the set of elementary events is infinite, then it is not obvious at all how one can "apply" the **Principle of Indifference** because even the formulation of it in the previous section looses its meaning: there is no permutation group in the infinite case with respect to which one could require invariance of the measure yielding the probabilities; equivalently: there is no uniform probability measure on an infinite set X of elementary events. It is therefore not clear at all what the **Principle of Indifference** *is* in connection with such infinite probability spaces and, consequently, even less obvious how one could apply the principle in such situations to obtain probabilities.

Having a look at the Bertrand's Paradox type arguments one can see however that in each of them the normalized restriction of the Lebesgue measure μ to compact subsets of $\mathbb{R}^n$ ($n = 1, 2, 3$) is used to express the Principle of Indifference of random events parametrized by elements of those compact sets: the probability of an event A is taken to be its Lebesgue measure $\mu(A)$. The Lebesgue measure on $\mathbb{R}^n$ is distinguished by the feature that it is the unique (up to multiplication by a constant) measure that is invariant with respect to the locally compact topological group of shifts in the Euclidean space $\mathbb{R}^n$. Such measures are called Haar measures. (Standard references for the Haar measure are Nachbin (1965) and Halmos (1950, Chap. XI.), for a more recent presentation see Deitmar and Echterhoff (2009).) This means, in effect, that in Bertrand's Paradox type arguments, the group of shifts is interpreted as expressing epistemic neutrality: this infinite topological group plays the role the permutation group plays in the finite case.

Generalizing, one can say that, given a topological group X the group action on X determined by X itself as a group can play the role of the action of the permutation group on X_n, and the Haar measure p_H on a compact X is the analogue of the uniform distribution on X_n if a non-zero uniform distribution on the elements X does not exist, which is the case if X is an infinite set. In what follows, $(X, \mathcal{S}, p_H)$ stands for a probability measure space in which X is a compact topological group with continuous group action, $\mathcal{S}$ is the Borel σ algebra on X and p_H is the Haar measure

on $\mathcal{S}$. These group and measure theoretic notions make it possible to formulate both the general classical interpretation of probability and the related principle of indifference *consistently* and generally as follows:

General Classical Interpretation: If X is a compact topological group, then the probabilities of the events are given by the Haar measure p_H on (the Borel sets of) X, and (**Frequency Link**:) the numbers $p_u(A)$ will be (approximately) equal to the relative frequency of A occurring in a series of trials producing elementary random events from X.

General Principle of Indifference: *If* X is a compact topological group and *if* the group action expresses epistemological indifference about the elementary random events in X, *then* the General Classical Interpretation is correct.

4 Labeling Invariance and Labeling Irrelevance

An intuition deeply ingrained in the elementary classical interpretation of probability is **Labeling Invariance**: the probability measure which is permutation invariant in one labeling assigns the same value of probability to the random events as the probability measure that is permutation invariant in a different labeling of the same random events: if one has a symmetric die sides of which are numbered by numbers $1, 2, \ldots 6$, then the permutation invariant measure on the elementary events is the uniform probability measure that assigns the value $\frac{1}{6}$ to each side. If, instead of labeling the sides by the numbers $1, 2, \ldots 6$, we label them by painting them using six different colours, then the measure which is invariant with respect to the permutation of colours will assign the value $\frac{1}{6}$ to each colour, the same value that is assigned to each side by the measure that is invariant with respect to the permutation of numbers. This is a trivial observation but has non-trivial consequences if it is interpreted erroneously. The wrong interpretation of this **Labeling Invariance** is that it is the proper expression of **Labeling *Irrelevance***: the intuition that labeling of random events is a matter of convention. **Labeling Irrelevance** is a very important condition in probabilistic modeling: its violation would entail a radical ambiguity and arbitrariness in assigning probabilities to random events. Violation of **Labeling Irrelevance** is obviously incompatible with any interpretation of probability that treats probability as an objective feature of the world. Note that **Labeling Invariance** is a sharply formulated, specific mathematical claim, whereas **Labeling Irrelevance** is a general and basic feature of application of probability theory to describe random phenomena conceived of as existing independently of any such modeling. Thus **Labeling Invariance** and **Labeling Irrelevance** are not directly comparable; however, **Labeling Invariance** can in principle be interpreted as the condition that ensures **Labeling Irrelevance**. Given the conceptual importance of **Labeling Irrelevance** and the (mistaken) identification of **Labeling Irrelevance** with **Labeling Invariance**, one expects **Labeling Invariance** to hold for the *general* classical interpretation as well. Does

it? In order to answer this question, we have to formulate **Labeling Invariance** precisely. To do this, one needs the notion of re-labeling first: If $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ are two probability spaces, then the map $h\colon X \to X'$ is called a re-labeling if it is a bijection between X and X' and both h and its inverse h^{-1} are measurable, i.e. it holds that

$$h[A] \in \mathcal{S}' \qquad \text{for all} \quad A \in \mathcal{S} \tag{2}$$

$$h^{-1}[B] \in \mathcal{S} \qquad \text{for all} \quad B \in \mathcal{S}' \tag{3}$$

(Here $h[A] = \{h(x) : x \in A\}$ and $h^{-1}[A'] = \{h^{-1}(x') : x' \in A'\}$.)

Labeling Invariance is the claim that the probabilities understood in the spirit of the general classical interpretation are invariant with respect to re-labeling: if $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ are two probability spaces (with p_H and p'_H being Haar measures) and h is a re-labeling between X and X', then the following holds:

$$p'_H(h[A]) = p_H(A) \qquad \text{for all} \quad A \in \mathcal{S} \tag{4}$$

$$p_H(h^{-1}[A']) = p'_H(A') \qquad \text{for all} \quad A' \in \mathcal{S}' \tag{5}$$

Recall (see e.g. Aaronson (1997, p. 3)) that two probability measure spaces $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ are called isomorphic if there are sets $Y \in \mathcal{S}$ and $Y' \in \mathcal{S}'$ such that $p(Y) = 0 = p'(Y')$ and there exists a bijection $f\colon (X \setminus Y) \to (X' \setminus Y')$ such that both f and its inverse f^{-1} are measurable and such that both f and f^{-1} preserve the measure p and p', respectively; i.e. (Eqs. 6–7) below hold:

$$p'(f[A]) = p(A) \qquad \text{for all} \quad A \in \mathcal{S} \tag{6}$$

$$p(f^{-1}[A']) = p'(A') \qquad \text{for all} \quad A' \in \mathcal{S}' \tag{7}$$

The function f is called then an isomorphism between the probability measure spaces. A compact expression of **Labeling Invariance** is therefore the following:

Labeling Invariance: Any re-labeling between probability spaces $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ is an isomorphism between these probability spaces.

5 General Bertrand's Paradox as Violation of Labeling Invariance

Labeling Invariance is a precise mathematical claim, and Bertrand's paradox can and should be interpreted as the statement that it does *not* hold in general:

Proposition 1 (General Bertrand Paradox, Gyenis and Rédei (2014)). *Let $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ be probability spaces with compact topological groups X and X' having an infinite number of elements and p_H, p'_H being the respective Haar measures on the Borel σ algebras $\mathcal{S}$ and $\mathcal{S}'$ of X and X'. Then **Labeling Invariance** does not hold for $(X, \mathcal{S}, p_H)$ and $(X', \mathcal{S}', p'_H)$ in the sense that*

- *either there is no re-labeling between X and X′;*
- *or, if there is a re-labeling between X and X′, then there also exists a re-labeling that violates* ***Labeling Invariance****.*

The **General Bertrand's Paradox** can be shown (Gyenis and Rédei 2014) to be a straightforward consequence of the following theorem in measure theory:

Proposition 2 (Rudin 1993; van Douwen 1984). *If X is an infinite, compact topological group with the Haar measure p_H on the Borel σ algebra $\mathcal{S}$ of X, then there exists an autohomeomorphism θ of X and an open set E in $\mathcal{S}$ such that $p_H(\theta[E]) \neq p_H(E)$.*

The General Bertrand's Paradox is thus a general feature of infinite probability measure spaces with the Haar measure yielding the probabilities. Note that the original Bertrand's Paradox only claimed that there exist Haar measures (namely the Lebesgue measure in $\mathbb{R}^n$) and re-labelings that violate **Labeling Invariance**: The General Bertrand's Paradox says much more: that *no two* infinite Haar probability spaces can satisfy **Labeling Invariance**. Thus Bertrand's 1888 Paradox is the specific "Lebesgue measure case" of a mathematical theorem on the Haar measure spaces defined by infinite topological groups that was proved in full generality only a hundred years after Bertrand's publication of the paradox.

It should now be clear what the possible strategies are if one wishes to defend **Labeling Invariance**: One can try to impose some extra conditions on re-labelings that entails either that re-labelings satisfying the extra conditions do not exist (Strategy A) or that the re-labelings satisfying the additional conditions force the re-labelings to be isomorphisms of the probability spaces (Strategy B). Bangu (2010) is an example of Strategy A. It can be shown that Bangu's suggestion for Strategy A is ambiguous however and that resolving the ambiguity makes it either a trivial case of Strategy B or is unsuccessful (see Gyenis and Rédei (2014) for details; for a different criticism of Bangu's attempt see Rowbottom and Schackel (2010)). A *successful* implementation of Strategy B would be to say that re-labelings are only admissible if they preserve our epistemic neutrality, i.e. if they are group isomorphisms. Clearly, these re-labelings do preserve the Haar measures. While this strategy is non-trivially successful in principle, all attempts to block the emergence of Bertrand's Paradox are misguided because Bertrand's Paradox (Proposition 1) is in complete harmony with probabilistic modeling as we argue next.

6 What Is the Problem with the Classical Interpretation of Probability?

Probability theory specified by the usual Kolmogorovian axioms is part of measure theory, hence it is pure mathematics, very much like geometry, for instance. But it can be used to describe certain phenomena. Such *applications* of probability theory should be distinguished from both pure mathematics and *interpretations* of probability.

In an application of probability theory one relates the mathematical elements in a given triplet $(X, \mathcal{S}, p)$ to specific non-mathematical entities. This involves two not unrelated yet conceptually different tasks:

Event Interpretation One has to specify what the elements in X and $\mathcal{S}$ stand for.
Truth Interpretation One has to clarify when the proposition "$p(A)$=r" is true/false.

In an application, probability theory thus becomes a mathematical *model* of a certain phenomenon that is external to mathematics. A probabilistic model is good if it has two features: it is descriptively accurate and predictively successful. Descriptive accuracy means that under the fixed specification of the Event and Truth Interpretations propositions such as $p(A) = r$ are true about events A that have been observed in the past. Predictive success means that the probabilistic propositions $p(A) = r$ will be true in future observations. Both descriptive correctness and predictive success are *empirical features*; furthermore, descriptive accuracy up until a given time T does *not* entail predictive success for times after T – this is just a particular formulation of the problem of induction.

We claim that the General Bertrand Paradox is in complete harmony with the application of probability theory as described above. To see this notice first that the mathematical notion of isomorphism between probability measure spaces also is in complete harmony with the application of probability theory: The Event Interpretation and Truth Interpretation are conceptually different issues, the former does not determine the latter, and, accordingly, two probability spaces are defined to be isomorphic if *two*, conceptually different conditions are satisfied: (i) the random events in the two spaces are connected by a re-labeling *and* (ii) the re-labeling preserves the probabilities. That is to say, a re-labeling is not necessarily an isomorphism in general, not even if the probability measures are of particular type: the General Bertrand's Paradox states that a re-labeling is indeed not a measure theoretic isomorphism in general in the category of Haar measure spaces with a group of infinite cardinality. From the perspective of the notion of isomorphisms of probability spaces *finite* probability spaces with the uniform probability measure just happen to have the feature that in this category re-labelings *are* isomorphisms.

The equivalence of measure theoretic isomorphism and existence of a re-labeling between finite probability spaces with uniform probability measures is however a contingent matter. But it can mislead one into thinking that **Labeling Invariance** is the proper way to ensure **Labeling Irrelevance**. It is precisely because of this conflation of **Labeling Invariance** and **Labeling Irrelevance** that violation of **Labeling Invariance** in the category of Haar measure spaces (i.e. General Bertrand's Paradox) appears paradoxical. But there is nothing paradoxical about this, **Labeling Irrelevance** is respected in probabilistic modeling perfectly well – *not* by **Labeling Invariance** but by the fact that measure theoretically isomorphic probability spaces describe the same random phenomenon: one is allowed to use any two pairs $(X, \mathcal{S})$ and $(X', \mathcal{S}')$ to label the same random events as long as there is a re-labeling h between X and X'. **Labeling Irrelevance** says that the choice of $(X, \mathcal{S})$ or $(X', \mathcal{S}')$ does not affect the probabilities of the random events. This is

indeed true, and is in complete harmony with the fact that choosing either $(X, \mathcal{S})$ or $(X', \mathcal{S}')$ does *not* determine any probability measure on either $(X, \mathcal{S})$ or $(X', \mathcal{S}')$: any of the (uncountably many) mathematically possible probability measures can be defined on both $(X, \mathcal{S})$ and $(X', \mathcal{S}')$. So, if the probability measure p is such that $(X, \mathcal{S}, p)$ is a descriptively accurate probabilistic model of the phenomenon in question, then the probability space $(X', \mathcal{S}', p')$ with p' defined by $p' \equiv p \circ h^{-1}$ also is a descriptively accurate model of the same phenomenon, and clearly then the measure spaces $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ are measure theoretically isomorphic with respect to h. Conversely, if $(X, \mathcal{S}, p)$ and $(X', \mathcal{S}', p')$ are isomorphic as probability spaces with respect to h, then (up to a measure zero set) h is a re-labeling of the events preserving probabilities, and either $(X, \mathcal{S}, p)$ or $(X', \mathcal{S}', p')$ can be used to describe the phenomenon, choosing any of them – choosing any of the two labelings $(X, \mathcal{S})$ or $(X', \mathcal{S}')$ in particular – is a matter of convention. It is important to realize that this interpretation of how **Labeling Irrelevance** is ensured in probabilistic modeling does not depend on any particular features of the probability spaces used in modeling; in particular it is not assumed that the probability measure is the uniform probability or a Haar measure. This is important and supports our description of how **Labeling Irrelevance** manifests in probabilistic modeling because it would of course be unacceptable if **Labeling Irrelevance** would only hold for situations in which the probabilities of the events are given by some special probability measures.

What are *interpretations* of probability? If one has a look at the philosophical literature about interpretations of probability one sees that interpretations are understood (more or less tacitly) as typical *classes* of applications of probability theory, classes of applications that share some common features, which the interpretation isolates and analyzes. The elementary classical interpretation concerns the application of the particular probability spaces $(X_n, \mathcal{S}, p_u)$, where the set X_n of elementary events is finite and the probability measure p_u is the uniform probability measure on X_n. The external-to-mathematics phenomena modeled by such probability spaces include flipping unbiased coins, throwing unloaded dice, card games – the classical paradigm examples of random phenomena. It should be clear now that the problem with the classical interpretation (understood with the amendment of the **Principle of Indifference**) is *not* Bertrand's Paradox, i.e. not that the **Principle of Indifference** cannot be consistently generalized from $(X_n, \mathcal{S}, p_u)$ to the case of an infinite number of elementary random events: We have seen in Sect. 3 that a **General Principle of Indifference** and a **General Classical Interpretation** *can* be consistently formulated in terms of Haar measure spaces $(X, \mathcal{S}, p_H)$ and one *can* in principle maintain the **General Classical Interpretation**. What is not maintainable is the general version of **Labeling Invariance**, which one has to give up as a consequence of the General Bertrand Paradox. It should also be clear now that the **Principle of Indifference** and **Labeling Invariance** are *independent*: One can abandon **Labeling Invariance** without violating the conceptually important **Labeling Irrelevance** because **Labeling Invariance** is different from **Labeling Irrelevance**.

All this should not be regarded as a defence of the classical interpretation of probability. The classical interpretation is mistaken. The main problem with it is simple and deep, and it is conceptual-philosophical: The classical interpretation misunderstands the basic nature of application of probability theory; specifically it disregards the empirical character of the applications of probability theory and gives the impression that descriptive accuracy and predictive success of applications of probability theory can be ensured by referring to an priori-flavored principle that expresses some sort of epistemic indifference about random events. But this is not possible, which is shown clearly also by the difficulty (often pointed out in connection with the **Principle of Indifference** (Hájek 2012)) that it is not clear how to specify the content of "epistemic neutrality" in such a way that the **Principle of Indifference** does not become circular and holds nevertheless: The **Principle of Indifference** holds only if epistemic neutrality *does* entail that the probabilities of the events given by the uniform probability measure *will* be equal to the frequencies of events in actual trials producing elementary random events, and such a conclusion cannot be validly based on a priori considerations – if it could, the **Principle of Indifference** would have solved the problem of induction.

One might say that the classical interpretation and the **Principle of Indifference** should be taken not with the **Frequency Link** but with the **Degree of Belief Link**, according to which p_H should be viewed as representing degrees of belief (Castell 1998; Mikkelson 2004). To assess the viability of such an interpretation of the classical interpretation one has to distinguish two further specifications of the notion of degree of belief: *descriptive* and *normative*.

In the descriptive interpretation the claim is that in some applications p_H represents the degree of belief of a particular person (or a specific group of people) about random events happening if the persons are epistemologically neutral about the events. Whatever the precise content of this epistemological neutrality, this descriptive interpretation of the degrees of belief is a claim about the thinking and behavior of certain empirically observable people. This claim is then in principle empirically testable, and it may or may not be true about people observed in the past and it may or may not be true about people to be observed in the future. Testing all this (including testing if the people in question do indeed have degrees of belief describable by a probability measure) is a matter for empirical psychology.

In the normative interpretation p_H is declared to stand for the *rational* degrees of belief of an abstract person (agent) if the agent is epistemologically neutral about the elementary events. In this case one has to ask further in what sense and why p_H represents *rational* degrees of belief? One answer can be that p_H is rational if $(X, \mathcal{S}, p_H)$ is a good model of a certain phenomenon in the sense described earlier in this section, and a rational agent's belief better be in harmony with the probabilities provided by a good model. This interpretation of rationality of p_H is essentially the content of the Principal Principle (Lewis (1986), Gyenis and Rédei (2013) and the references therein) and, while it is very natural, one should realize that the probability measure p_H features in it in *two*, distinct roles: (i) standing for the degree of belief *and* (ii) representing some extra-mental, non-degree-of-belief-type quantities (for instance "chances", frequencies or some other dimensionless

physical quantities (Szabó 2007)) with which the degrees of belief are required to be equal. Thus this interpretation reduces the **Degree of Belief Link** to another Interpretive Link and thereby the rationality (or otherwise) of an agent's degree of belief is made again dependent on empirical matters. But then it does not matter from the perspective of rationality of the degrees of belief whether the agent is epistemically neutral about the elementary events or not, because the correctness of the probabilistic model is an empirical matter that cannot be ensured on the basis of an a priori neutrality, and probability measures different from p_H can very well be rational if they satisfy the Principal Principle and the probabilistic model is good. Another possible specification of rationality of the agent's degrees of belief can be that they are consistent, i.e. that p_H satisfies the axioms of probability. Obviously, this does not single out p_H as the only rational probability. In sum: the **Degree of Belief link** ultimately reduces the content of validity of the Principle of Indifference either to matters of fact studied in empirical psychology or to logical consistency, which is too weak to establish the General Classical Interpretation.

References

Aaronson, J. (1997). *An introduction to infinite Ergodic theory* (Volume 50 of mathematical surveys and monographs). Providence: American Mathematical Society.

Aerts, D., & de Bianchi, M. S. (2014). Solving the hard problem of Bertrand's Paradox. *Journal of Mathematical Physics, 55*. Published online 21 July 2014. doi:10.1063/1.4890291.

Bangu, S. (2010). On Bertrand's Paradox. *Analysis, 70*, 30–35.

Bertrand, J. L. F. (1888). *Calcul de Probabilités*. Paris: Gauthier-Vilars.

Castell, P. (1998). A consistent restriction of the Principle of Indifference. *The British Journal for the Philosophy of Science, 49*, 387–395.

Deitmar, A., & Echterhoff, S. (2009). *Principles of harmonic analysis* (Universitext). New York: Springer.

Gyenis, Z., & Rédei, M. (2013). Can Bayesian agents always be rational? A principled analysis of consistency of an Abstract Principal Principle. Submitted. http://philsci-archive.pitt.edu/10085/.

Gyenis, Z., & Rédei, M. (2014). Defusing Bertrand's Paradox. *The British Journal for the Philosophy of Science*. Forthcoming, online: 28 Mar 2014. doi:10.1093/bjps/axt036.

Hájek, A. (2012). Interpretations of probability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2012 edition)*. http://plato.stanford.edu/archives/sum2012/entries/probability-interpret/. Accessed 29 May 2012.

Halmos, P. (1950). *Measure theory*. New York: D. Van Nostrand.

Lewis, D. (1986). A subjectivist's guide to objective chance. In *Philosophical papers, vol. II* (pp. 83–132). Oxford: Oxford University Press.

Marinoff, L. (1994). A resolution of Bertrand's Paradox. *Philosophy of Science, 61*, 1–24.

Mikkelson, J. M. (2004). A resolution of the wine/water paradox. *The British Journal for the Philosophy of Science, 55*, 137–145.

Nachbin, L. (1965). *The Haar integral*. Princeton: D. Van Nostrand.

Rowbottom, D. (2013). Bertrand's Paradox revisited: Why Bertrand's 'solutions' are all inapplicable. *Philosophia Mathematica, 21*, 110–114.

Rowbottom, D. W., & Schackel, N. (2010). Bangu's random thoughts on Bertrand's Paradox. *Analysis, 70*, 689–692.

Rudin, W. (1993). Autohomeomorphisms of compact groups. *Topology and Its Applications, 52*, 69–70.

Shackel, N. (2007). Bertrand's Paradox and the principle of indifference. *Philosophy of Science, 74*, 150–175.

Szabó, L. E. (2007). Objective probablity-like things with and without objective indeterminism. *Studies in the History and Philosophy of Modern Physics, 38*, 626–634.

van Douwen, E. K. (1984). A compact space with a measure that knows which sets are homeomorphic. *Advances in Mathematics, 52*, 1–33.

von Mises, R. (1928). *Probability, statistics and truth* (2nd ed., 1981). New York: Dover. Originally published as 'Wahrscheinlichkeit, Statistik und Wahrheit'. Springer, 1928.

Revisiting Smale's Fourteenth Problem to Discover Two Definitions of Chaos

L.C. Zuchowski

1 Introduction

Smale (1998, p. 14) proposes the following question as one of eighteen mathematical problems to be solved in the twenty-first century:

> Is the dynamics of the ordinary differential equations of Lorenz that of the geometric Lorenz attractor of Williams, Guckenheimer, and Yorke?

Later in the article, he clarifies that the question aims to establish whether the system of Lorenz equations is chaotic in the same sense as the horseshoe map he himself investigated in Smale (1967). Tucker (2002) has shown that rigorously constructed numerical solutions to Lorenz's system support an attractor similar to the analytic one. For many, he has thereby solved Smale's fourteenth problem.

In this paper, I revisit Smale's fourteenth problem and address the question whether the definition of chaos that has been applied to Smale's horseshoe map is the one that also fits the Lorenz equations. For philosophy of science, the question is less one of mere partial formal equivalence (as addressed by Tucker 2002) but one of the construction and use of definitions in mathematics. Therefore, while I aim to offer a 'sideways' look at Smale's fourteenth problem, the main contribution of this paper might well be to the as yet unresolved debate about the definition of chaos (e.g. Kellert 1993; Werndl 2009). As such, my discussion will be based neither on an in-depth analysis of the Lorenz equation nor on one of Smale's horseshoe map. Instead, I will look at a simpler iconic object of chaos, the logistic equation.

In Sect. 2, I will show that there exist three different ways of analysing the logistic equation: through analytic integration (Sect. 2.1); through successive

L.C. Zuchowski (✉)
Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK
e-mail: lena.zuchowski@sbg.ac.at

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_21

self-application (Sect. 2.2) and through numerical integrations (Sect. 2.3). In Sect. 2.4, I will show that these different 'takes' on the logistic equation reveal very different properties.

Furthermore, in Sect. 3, I will argue that two different definitions of chaos can be justified by these different kinds of analysis. One of these, name Periodic Chaos (Sect. 3.1), includes the Smale horseshoe map as its natural-world justification; the other, named Aperiodic Chaos (Sect. 3.2), was developed to captured the properties of systems like the Lorenz equations. I will also argue that the coexistence of these two definitions of chaos is unarticulated in the existing debate – but that it is not a sign of conceptual inconsistency since both are justified through different criteria.

In conclusion (Sect. 4), my answer to Smale's fourteenth problem is a cautious negative one: it appears that the Lorenz equations and the horseshoe map exemplify rather different kinds of chaos and that the most promising answer to the riddle might be in the recognition of these differences.

2 Three Takes on the Logistic Equation

The name "logistic equation" is given to the following non-linear, first-order differential equation (e.g. Devaney 1989; Hilborn 2002):

$$\frac{dN}{dx} = rN\,(1 - N)\,, \tag{1}$$

where N is the dependent variable, x is the independent variable, and r is a constant. Since the logistic equation has most often been used in population dynamics, x is sometimes (but, as we will see below, not always) taken to be a time variable and N is a measure of the population number. The constant r is called the growth rate.

The logistic equation has become a paramount example of chaos (e.g. Devaney 1989; Hilborn 2002). However, it is my aim to show that there are actually three different prevalent ways of analysing the equation: analytic integration (Sect. 2.1); successive self-application (Sect. 2.2); and numerical integration (Sect. 2.3). These three kinds of analysis yield very different qualitative and quantitative results. In particular, only the latter two ways of handling (1) produce expressions for N that have been called 'chaotic'. I will also show that a different set of features results from numerical integration than from successive self-application. In Sect. 2.4, I will outline the most important differences between the three approaches to the logistic equation and also show how a disregard for these formal differences has led to confusion in the literature on the logistic equation.

In Sect. 3, we will then see that this implies that different definitions of chaos are related to these two latter kinds of analysis.

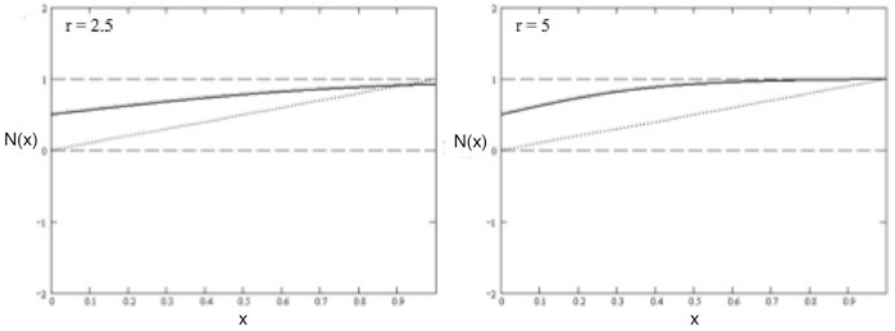


Fig. 1 The function $N(x)$ resulting from the analytic integration of the logistic equation (*solid line*) in comparison to the identity function (*dotted line*). The functions shown here use $N_0 = \frac{1}{2}$

2.1 Take 1: Analytic Integration

Equation (1) can be straightforwardly analytically integrated (e.g. by separation of variables):

$$N(t) = \frac{N_0 e^{rx}}{1 + N_0 (e^{rx} - 1)}, \tag{2}$$

where N_0 is an initial value for $N(x)$. It is immediately apparent that, independent of the initial value N_0, for $x \to \infty$, $N(x)$ approaches 1 for all positive values of r and approaches zero for all negative values of r. Here, we will only deal with positive r values.

Figure 1 shows the function $N(x)$ for two different values of r, $r = 2.5$ and $r = 5$. The qualitative behaviour of the function is the same for all r values: it monotonically approaches one. The larger the r value, the more rapidly this approach occurs. It is also apparent from Fig. 1 that, independent of our choice for r, the function only intersects the identity function once. There is therefore only one fixed point of the form $N(x) = x$. It is also easy to verify that the above characteristics are independent of our choice for the initial value N_0.

2.2 Take 2: Successive Self-Application

The second take on the logistic equation is based on the discussion given by Devaney (1989). My own description in this section is based on his analysis but does not aim to provide formal justification or mathematical comprehensiveness: instead, the reader is referred to the original source to look up formal proofs and derivations. Whenever possible, I will refer to Figs. 2 and 3 to illustrate the claims made in this section.

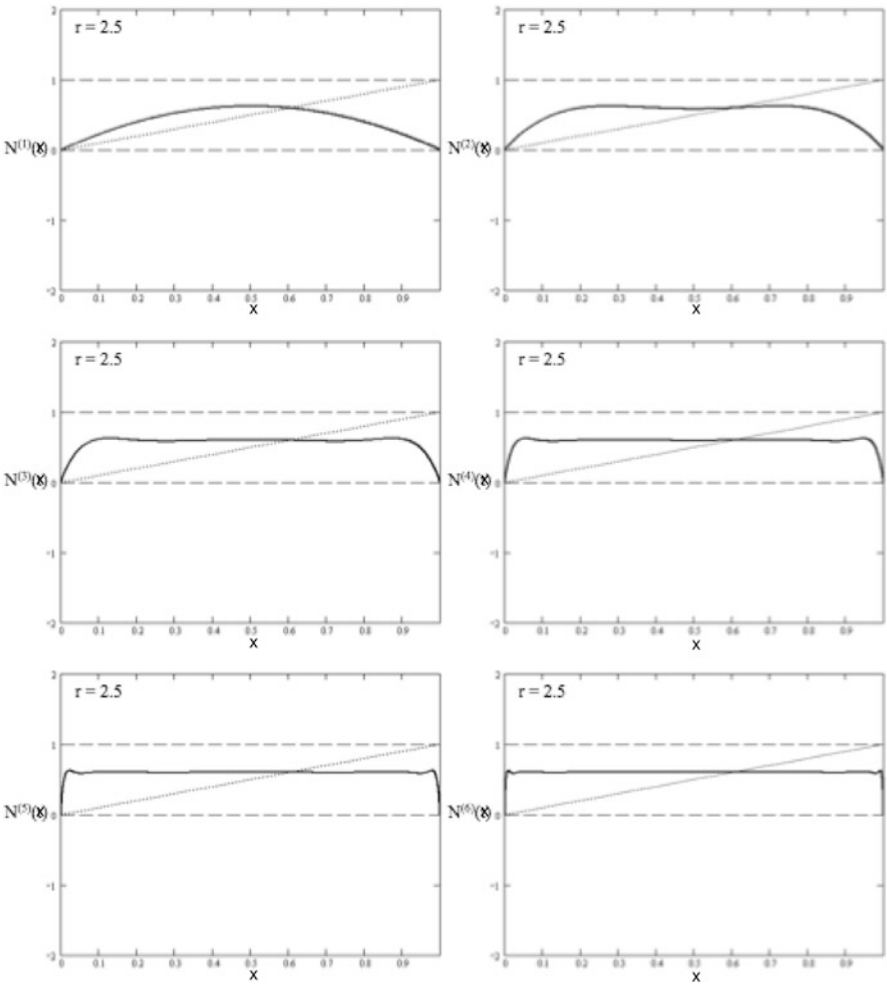


Fig. 2 The functions $N^{(n)}(x)$ resulting from subsequent self-application of the logistic equation with growth rate $r = 2.5$ (*solid line*) in comparison to the identity function (*dotted line*)

Successive self-application Successive self-application means that the right-hand side of (1) is successively applied to itself:

$$\begin{aligned} N^{(0)}(x) &= x, \\ N^{(1)}(x) &= rx(1-x), \\ N^{(2)}(x) &= rN^{(1)}(1-N^{(1)}) \\ &\cdots \\ N^{(n+1)}(x) &= rN^{(n)}(1-N^{(n)}). \end{aligned} \tag{3}$$

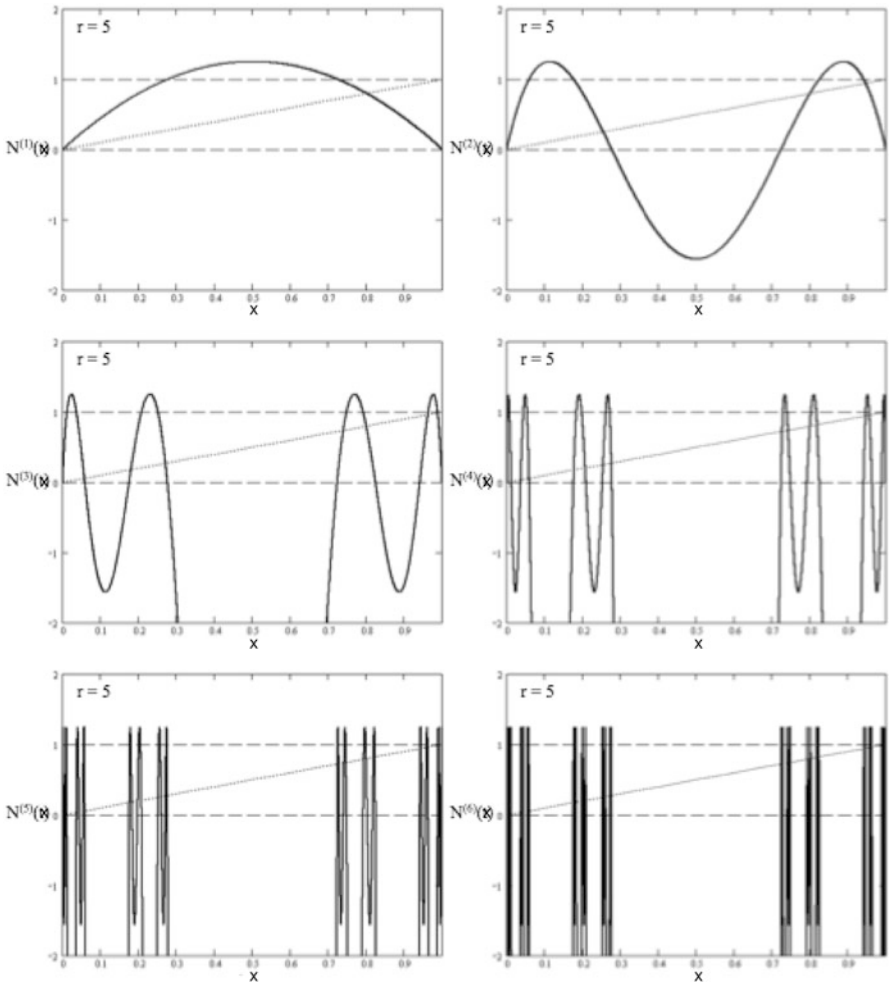


Fig. 3 The functions $N^{(n)}(x)$ resulting from subsequent self-application of the logistic equation with growth rate $r = 5$ (*solid line*) in comparison to the identity function (*dotted line*)

It is easy to see that this will result in subsequently higher polynomial terms in x:

$$\begin{aligned} N^{(0)}(x) &= x, \\ N^{(1)}(x) &= rx - rx^2, \\ N^{(2)}(x) &= r^2x - (r^2 + r^3)x^2 + 2r^3x^3 - r^3x^4 \\ &\dots \end{aligned}$$

The order of the polynomial in x will be n^n for even n and $(n-1)^{2(n-1)}$ for odd n. The order of the polynomial in r will be $(n-1)^{n-1}+n$.

Devaney (1989) notes that the growth rate r crucially influences the behaviour of the function $N(t)$ under self-iteration. In particular, for $r > 4$, part of the function will always be mapped outside the initial domain, i.e. the unit interval. This is clearly visible in Fig. 3. In contrast, for $r \leq 4$, the function will always map the unit interval onto itself, as illustrated in Fig. 2. This difference results in very different qualitative behaviours in the two parameter regimes.

Asymptotic behaviour for $r < 3$: For $r < 3$, the behaviour is simple and mimics that of the analytically integrated function (2). As shown in Fig. 2 successive self-iteration leads to an asymptotic approach of the function to a constant equilibrium value:

$$N_{eq} = \frac{r-1}{r}. \tag{4}$$

The equilibrium value is also the value of the function's single, non-trivial fixed point. The behaviour in the parameter regime $r > 4$ is much more complicated and will be described below.

Transitioning to infinitely many periodic points: Devaney (1989, pp. 60–67) discusses the implications of the fact that for $r = 3.839$, (3) has a periodic point of period 3. According to Sarkovskii's theorem, this implies that the map also has periodic points of all other periods. In detail, Sarkovskii's theorem reads, according to Devaney (1989, p. 60):

Theorem 1. *Let $f : \mathbf{R} \to \mathbf{R}$ be continuous. Suppose f has a periodic point of period 3. Then f has periodic points of all other periods.*

The parametric region $3 \leq r \leq 4$ is thus characterized by a transition from asymptotic behaviour to a dynamics characterized by infinitely many periodic points. Since the majority of these periodic points will be unstable, they will not be visible in computer-generated graphics.

Removing the middle for $r > 4$: For $r > 4$, we know that the maximum value of $N^{(1)}$ is $\frac{r}{4}$ and hence is greater than 1. This implies that there exists a set of values A_0 which are mapped outside the unit interval under the first iteration. Looking at Fig. 3, we can see that A_0 is an open interval centred around $x = \frac{1}{2}$. Devaney (1989) then shows that points that have been mapped outside the unit interval once, will be mapped monotonically to $-\infty$ through subsequent self-applications. For A_0, this is well illustrated in Fig. 4. A different, more memorable, way of describing this behaviour is to say that points which have "escaped" (p. 34) the unit interval once, have escaped the interval for good.

It remains for us to consider the fate of the points staying in the unit interval. Following Devaney's convention, we will denote the set of points doing so up to an iteration n as:

$$\Lambda_n := I - \left(\cup_{k=0}^{n} A_k\right),$$

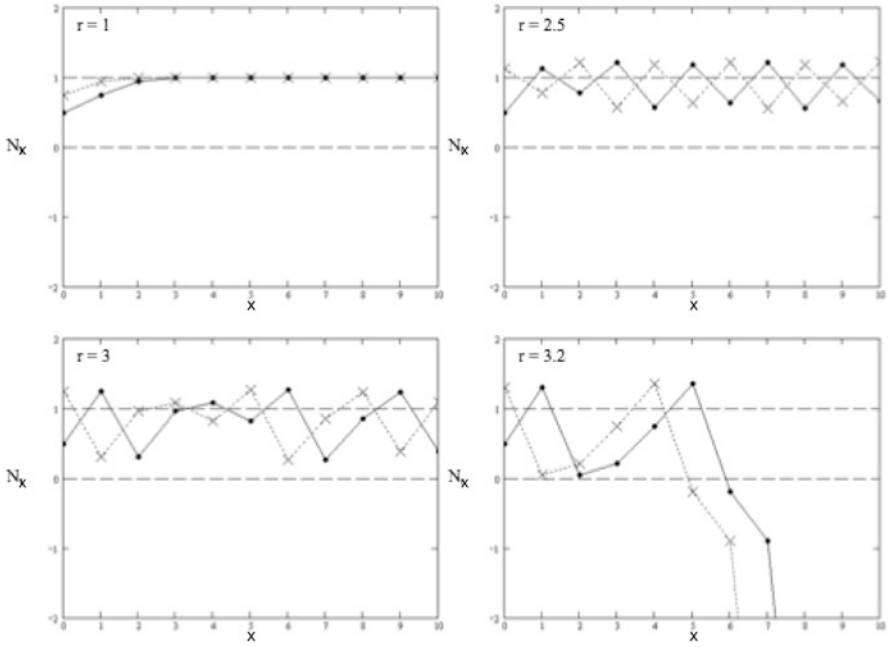


Fig. 4 Population values N_x (*dots*) in comparison to the values of the next iteration N_{x+1} (*crosses*). Those values where the dots and crosses coincide are fixed points. The iterations shown here use $N_0 = \frac{1}{2}$

where A_k is the set of points escaping during the k-th iteration. Each Λ_n is composed of the $2^{n+1} - 1$ closed intervals lying in between the open intervals of escaping points. The set of points remaining forever within the unit interval is therefore given by:

$$\Lambda = I - \left(\cup_{k=0}^{\infty} A_n\right), \tag{5}$$

It is intuitively graspable that the construction above could be described as 'removing the middle'. Devaney (1989, pp. 37–38) shows that it is indeed formally equivalent to the construction of a Cantor Middle Third set and that the resulting set Λ will be a Cantor set. This also implies that Λ will be fractal, i.e. self-similar under magnification.

Fixed points in Λ: Since all of the escaping points A_n are subsequently mapped to $-\infty$, the graph needs to have an intermediate maximum of subsequently escaping points on each of the closed intervals forming Λ_n and hence form a 'hump' on each of those intervals. Accordingly, the nth-iteration of (3) crosses the line $N(x) = x$ at least 2^n times and hence has at least a number $P_n(N) = 2^n$ of fixed points. Accordingly, one also finds:

$$P_n(N) \to \infty, \quad for \ \ n \to \infty. \tag{6}$$

A further characteristic that Devaney (1989) proves for the logistic equation is also immediately apparent from Fig. 3: the set of fixed points is dense in Λ. Here we recall:

Definition 1. A subset U of S is dense in S if S is the closure of U, i.e. S consists of U and all its limit points.

Roughly speaking, under an infinite iteration of the 'removing the middle' procedure described above, the disjoint sets forming Λ will become exceedingly small. However, since they are forced to each contain a fixed point, each such subset will eventually consist of this fixed point and its limit points.

Topological transitivity: Devaney (1989, pp. 47–50) uses a technically sophisticated proof (relying on symbolic dynamics and the topological conjugacy of the logistic map on Λ, for $r \geq 2 + \sqrt{5}$, to a shift map) to show that N is topologically transitive on Λ. We recall:

Definition 2. A map $f : J \to J$ is topological [sic] transitive if for any pair of open sets $U, V \subset J$, there exists $n > 0$ such that $f^{(n)}(U) \bigcap V \neq \emptyset$.

This implies that, eventually, points from one arbitrarily small neighbourhood will be mapped into any other arbitrarily small neighbourhood.

Sensitivity to initial conditions: In a similar procedure relying on the conjugacy to the shift map, Devaney (1989, p. 49) also shows that the logistic map on Λ possesses sensitive dependence on initial conditions for $r \geq 2 + \sqrt{5}$, in the sense that:

Definition 3. $f : J \to J$ has sensitive dependence on initial conditions if there exists $\delta > 0$ such that, for any $x \epsilon J$ and any neighborhood X of x, there exists $y \epsilon X$ and $n \geq 0$ such that $|f^{(n)}(x) - f^{(n)}(y)| > \delta$.

Roughly speaking, this means that there is at least one point arbitrarily close to x which will eventually be separated from it by at least δ. Defining the distance δ requires the existence of a metric; but in our example it is the elementary one on the number line. Figure 3 illustrates the existence of sensitive dependence on initial conditions in the sense that the continued narrowing of the intervals in Λ_n will lead to more and more points in the neighbourhood of a fixed point being eventually mapped to $-\infty$. Hence any neighbourhood around an $x \epsilon \Lambda$, which contains a part of A_n, can be used to demonstrated sensitivity to initial conditions.

Applicability: It is notable that Devaney (1989) initially only provides proofs of these last two properties for a growth rate $r > 2 + \sqrt{5}$ and only on the set Λ, which covers only a small fraction of the unit interval. Further proofs are provided for selected other cases: for example, Devaney (1989, pp. 50–51) shows that all three properties discussed above hold for $N(x) = 4x(1 - x)$ on the whole unit interval; further proofs for particular logistic functions are set as exercises for the reader.

2.3 *Take 3: Numerical Integration*

A third take on the logistic equation is provided by such authors as May (1974) and May and Osler (1976). Integrating (1) numerically one obtains:

$$\begin{aligned} dN &= rN\,(1-N)\,dx \\ N_{x+1} - N_x &= rN_x\,(1-N_x)\,\Delta x, \end{aligned} \tag{7}$$

where the subscript x denotes the x-th iteration and Δx is the difference between two such steps of the independent variable. If the dynamics described is inherently discrete, e.g. if N_x describes the size of a generation, then we can assume $\Delta x = 1$ and hence:

$$N_{x+1} = N_x\,(1 + r\,(1-N_x))\,. \tag{8}$$

It is notable that x here is a discrete variable rather than a continuous one as in Sects. 2.1 and 2.3. Equation (8) is also called the "logistic distance equation".

Equation (8) is one of the equations studied by May (1974) and May and Osler (1976). My discussion below will roughly follow their treatment of the logistic equation. Again, I do not aim to present a formal mathematical analysis and, whenever possible, will illustrate properties by referring to Fig. 4 rather than proving their existence formally.

Asymptotic behaviour for $r < 2.5$: The development of the iterations for growth rates below $r = 2.5$ is illustrated in the first panel of Fig. 4. As can be seen, the values on N_x will converge to 1, which eventually becomes a fixed point of the map. The speed of this convergence is proportional to the growth rate r. The behaviour of the numerically integrated function thus closely mirrors that of the analytically integrated one.

Transitioning to infinitely many periodic points for $2.5 \leq r \leq 2.7$: For growth rates between $r = 2.5$ and $r = 2.7$, the iterated values behave periodically, i.e. some values N_i will repeat themselves at N_{i+p}, N_{i+2p} ..., where p is the period of the repetition. This parameter regime is also illustrated in Fig. 4. May and Osler (1976) attribute the increase in periodicity to Sarkovskii's theorem (cf. Theorem 1). However, the proof of the theorem given in Appendix A of that paper is a general one for continuous maps.

Aperiodicity for $r > 2.7$: About the behaviour of the function beyond $r = 2.7$, May and Osler (1976, p. 583) write:

> [T]here are an infinite number of periodic points. ... Furthermore, there are an unaccountable [sic] number of points (initial conditions) whose trajectories are totally aperiodic; no matter how long the time series generated by [the difference equation] is run out, the pattern never repeats [].

The aperiodic behaviour of the series of points N_x is illustrated in the two relevant cases displayed in Fig. 4. The general idea behind this is that these are the iterated sets of points ($N_0, N_1, N_2, \ldots, N_\infty$) created by (8) which do not contain a fixed point of $N^{(p)}(x = N_x)$ or any value of p. Li and Yorke (1975, p. 987) show that there are an uncountable number of initial values N_0 for which such series will be created. They were also the first to use the term chaos to describe these sets of points. Both May and Osler (1976, pp. 583–585) and Guckenheimer et al. (1977, p. 105) remark that aperiodic sets will be undistinguishable from periodic series with a period k large enough to fall outside any reasonable iteration interval. These points would then for practical purposes behave aperiodically.

Extinction for $r > 3$: A fourth regime is identified by May and Osler (1976, pp. 586–588): for growth rates higher than $r = 3$, the iterated population values will increasingly fall below zero. This is described as 'extinction'. The last panel of Fig. 4 illustrates the behaviour of the sequence in this regime.

2.4 Comparison and Confusion

My description of the important differences between the three takes on the logistic equation will focus on three main points: the use of analytic or iterative modes of analysis; the use of a continuous or discrete independent variable x; and the diagnosis of aperiodicity or infinity periodicity. The characteristic of the three takes on the logistic equation with respect to these categories are summarized in Table 1. I will also describe how some of these differences are habitually obscured in the literature on the logistic equation.

Analytic or iterative mode of analysis: It is immediately apparent from the descriptions in Sects. 2.1, 2.2 and 2.3 that only the integration shown in Sect. 2.1 (i.e. Take 1) leads to an analytic expression. Take 2 and take 3 employ iterative modes of analysis in the form of successive self-application and numerical integration, respectively.

The mode of analysis influences the behaviour observed: the analytic expression (2) displays only asymptotic development while the two iterative expressions, (4) and (8), possess more 'interesting' properties. However, these differences are only apparent for large growth rates r; for small growth rates all three cases behave asymptotically. Notably, it is only in the latter case that analytic and iterative modes of analysis lead to similar results.

Table 1 Comparison of the different approaches to the logistic equation outlined in Sects. 2.1, 2.2 and 2.3

	Mode of analysis	Independent variable	Aperiodicity
Take 1	Analytic	Continuous	No
Take 2	Iterative	Continuous	No
Take 3	Iterative	Discrete	Yes

Continuous or discrete independent variable: It is easy to see that the analysis in Sect. 2.1 uses both a continuous independent variable x as well as a continuous dependent variable $N(x)$.

Despite the fact that the mode of analysis is iterative in Sect. 2.2, analysis through self-application also uses both a continuous independent variable x as well as a continuous variable $N(x)$. The fact that Fig. 3 appears to show a discrete function for higher iterations of the self-application algorithm is only a consequence of the limiting resolving power of the graphics program and the restricted plotting domain. Admittedly, much of the discussion in Sect. 2.2 takes place on the level of sets of points. However, these sets are fundamentally derived as intervals of the domain of a continuous map; they become fully discrete only in the limit of infinitely many self-applications.

In contrast to Sects. 2.1 and 2.2, the numerical integration in Sect. 2.3 uses a discrete independent variable x. It also works on a discrete real value N_x and maps it into another discrete real value N_{x+1}. The values are labeled by the discrete parameter x, which – in contrast to the x values in Sects. 2.1 and 2.2 – can only assume integer values. The discreteness of (8) is often obscured in discussions of the logistic equation by plotting the connecting line between the values and not the discrete values themselves (e.g. May 1986, Figure 5). This practice is particularly prevalent if plots of the real dependent variable N_x against N_{x+1} are presented (e.g. Hilborn 2002, Figure 1.10; May and Osler 1976, Figure 4). Assuming that we are using the standard topology on the number line, such a representation ignores the fact that both x and $N(x)$ can only take on discrete values (albeit only the former is restricted to integers) if we wish to work with (8): to get a continuous spread of N_x values, one would simultaneously have to decrease the step size Δx in (7). However, to do so would crucially alter the nature of the equation and the behaviour displayed. Agreed, an equivalence of the set N_x could conceivably be established by the choice of an appropriate, non-standard topology. However, none of the sources discussed here gives any indication that the topology on their number space has been designed to do so.

Diagnosis of aperiodicity: In the two cases that display regime-dependent, non-asymptotic behaviours, only take 3 (Sect. 2.3) gives a diagnosis of aperiodicity. Since both behaviours are formally linked by Sarkovskii's theorem, there is no formal contradiction. However, there is a significant difference is emphasis, since aperiodicity was not even mentioned in Sect. 2.2. As we will see in Sect. 3, this has far-reaching consequences for the definition of chaos.

3 Two Kinds of Chaos

Building on the description of the three takes on the logistic equation identified in Sect. 2, I will now show that different definitions of chaos have been associated with Take 2 (Sect. 2.2) and Take 3 (Sect. 2.3). In particular, I will show that the behaviour of the logistic equation under self-application has served as iconic phenomenon for the definition of *periodic chaos* (Sect. 3.1) while the behaviour of the logistic

equation under numerical iteration has served as a iconic case for the definition of *aperiodic chaos* (Sect. 3.2). As indicated by the monikers, the former puts crucial emphasis on infinite periodicity while the latter stresses aperiodicity.

To do so, we will a framework for evaluating these various definitions comparatively. I will adopt the one used by Werndl (2009), who advocates three possible criteria for the justification of definitions in mathematics:

(i) **Natural-world justification:** The definition captures a pre-formal idea about the natural world.
(ii) **Condition justification:** The definition corresponds to a valuable mathematical idea.
(iii) **Redundancy justification:** The definition eliminates a redundant idea in an existing definition.

In this paper, I am primarily concerned with the conception of what I consider two base-line definitions of chaos and will therefore focus on criteria (i) and (ii). I will show in Sects. 3.1 and 3.2 that the fact that Periodic Chaos and Aperiodic Chaos are based on different natural-world phenomena – in the first instance, behaviour of the logistic equation according to Take 2 and Take 3, respectively – and correspond to different mathematical concepts – most prominently, infinity periodicity and aperiodicity, respectively – means that *both* of them are justified according to criteria (ii) and (ii). The justification characteristics of the two definitions are summarized in Table 2.

It is important to note here that the justification criteria are based on the idea that these are the properties that make a definition acceptable to the mathematical community rather than statements about the ontology of the concept (Werndl 2009). My analysis in Sect. 2 also raises questions about the ontology of chaos. However, given the constraints on the scope of this paper, we will set these aside for the moment.

3.1 Periodic Chaos

Devaney's (1989, p. 50) own formal definition of chaos, which specifies three criteria to be fulfilled simultaneously by a chaotic map f, is the following:

Table 2 Justification of Periodic Chaos and Aperiodic Chaos according to Werndl's (2009) criteria (i) and (ii)

	(i) Natural-world idea		(ii) Math. concept
Periodic Chaos	**Log. Equ.** Self-application	**Others** Horseshoe maps Billiard dynamics (Fractals)	Infinite periodicity transitivity (SDIC)
Aperiodic Chaos	Numerical integration	Numerical solutions Experimental data	Aperiodicity (SDIC)

Definition 4. Let V be a set. $f : V \to V$ is said to be chaotic if

1. f has sensitive dependence on initial conditions; and
2. f is topologically transitive; and
3. periodic points are dense in V.

We have shown in Sect. 2.2 that the logistic equation under successive self-application fulfills all three criteria and is therefore chaotic.

Natural-world justification: In fact, the discussion in Devaney (1989, pp. 31–53) leaves no doubt that Definition 4 has been developed specifically to describe the properties of (3) as revealed in Take 2. We can thus view the features revealed under Take 2 on the logistic equation as part of a natural-world justification of Definition 4. There are a number of other one-dimensional maps that do so (under successive self-application), including $N(x) = \tan x$; the baker map and the tent map (Devaney 1989, p. 52). In the second part of the book, Devaney (1989, pp. 157 ff.) also shows that there are higher dimensional maps that are chaotic according to Definition 4, most importantly all those that are similar to the horseshoe map discovered by Smale (1967).

Moving away from the idea of successive self-application, Berger (2001) cites Definition 4 in a chapter devoted to describing the dynamics of various billiard situations.

Another phenomenon that can be counted as part of the natural-world justification of Definition 4 are fractals. Fractal sets are not maps and therefore not as such covered by Definition 4. However, as described in Sect. 2.2, the periodic points of chaotic maps often form fractal sets. I am therefore inclined to count fractals as at least auxiliary parts of the natural-world justification of the definition. The natural-world justification for Definition 4 are listed in Table 2. The list does not claim to be comprehensive. However, it appears to me to cover the majority of phenomena discussed in the relevant literature.

Condition justification: The crucial mathematical concepts associated with Definition 4 are transitivity (criterion 2) and dense periodicity (criterion 3). Criterion 1, sensitivity to initial conditions has been shown to be a consequence of the latter two criteria (Banks et al. 1992). I have still included it into the list of condition justifications since it is often perceived as the most salient characteristic of chaos (e.g. Ruelle 1991). All these justifications have been included in Table 2. The most striking one and also the one that Devaney (1989) places most emphasis on (cf. Sect. 2.2) is clearly the dense (infinite in most cases). Accordingly, I have named this kind of behaviour periodic chaos.

3.2 Aperiodic Chaos

May (1974, 1986) and May and Osler (1976) – which I have chosen as prototype studies using Take 3 on the logistic equation in Sect. 2.3 – appear to use the term

'chaotic' synonymously with 'aperiodic' in the sense discussed in that section. Equation (8) is therefore chaotic because it displays aperiodicity for certain values of r. Lorenz (1993, p. 4), whose Lorentz attractor also became part of the natural world justification of aperiodic chaos, provides a more formal definition:

Definition 5. I shall use the term *chaos* to refer collectively to processes of this sort – ones that appear to proceed according to chance even though their behaviour is in fact determined by precise laws.

The apparent stochasticity of chaotic processes is also stressed by May (1976, e.g. p. 583). As a consequence of this aperiodic behavior, the mathematical representations of such processes are also sensitive to initial conditions: small changes of the initial parameters will result in different aperiodic trajectories, which – by the very virtue of their aperiodicity – will be very dissimilar to each other (e.g. May and Osler 1976, p. 586; Lorenz 1993, p. 8).

Natural-world justification: From the examples discussed in Sect. 2, the behaviour providing a natural-world justification for Definition 5 is, of course, Take 3. It is interesting (but beyond the scope of this paper) that other, higher-dimensional systems providing such a justification are also the result of numerical integration: most prominently the famous Lorenz attractor already alluded to Lorenz (1963, 1993). Lorenz (1993) provides several other computer-generated examples, including a simply iterative algorithm of a skier in a mogul field. There now also exists a large and growing body of literature identifying aperiodically chaotic behavior in experimental data sets (e.g., for review, Kellert 1993, 2008).

Condition justification: The 'interesting' mathematical concept underlying Definition 5 is aperiodicity, in particular when created from relatively simple expressions like (8). This provides a sharp contrast to the notion of infinite periodicity we encountered in Sect. 3.1.

However, there are also similarities: as in Sect. 3.1 and Definition 4, an auxiliary justification is provided by the notion of sensitive dependence to initial conditions. However, while I have used the same descriptor for both these features, the notion of sensitivity to initial conditions in the definition of periodic chaos is formalized by Definition 3, while it is used in a much looser sense in the literature on aperiodic chaos. Some caution therefore needs to exercised when equating the two concepts.

4 Conclusion

My analysis in Sects. 3 and 4 shows that there are two different definitions of chaos, periodic chaos and aperiodic chaos. These two definitions emphasis different properties as essential for the existence of chaos. In particular, the former postulates infinite periodicity while the latter is based on aperiodicity.

Both definitions are justified according to the justification criteria put forward by Werndl (2009). Since the phenomena and notions drawn upon by each definition for

natural-world and condition justifications differ, this does not present a conceptual inconsistency. However, since Smale's horseshoe map is part of the justification for periodic chaos and the Lorenz equations are part of the justification for aperiodic chaos, we might be justified in concluding that Smale's fourteenth problem is of truly Gordian proportions; that is, rather than try to show that one is chaotic in the same sense as the other, we should cut the knot by simply recognizing that they are integral parts of different definitions of chaos.

References

Banks, J., Brooks, J., Cairns, G., Davis, G., & Stacey, P. (1992). On Devaney's definition of chaos. *The American Mathematical Monthly, 99*, 332–334.

Berger, A. (2001). *Chaos and chance: An introduction to stochastic aspects of dynamics*. New York: Walter de Gruyer.

Devaney, R. L. (1989). *An introduction to chaotic dynamical systems*. Redwood City: Addison Wesley.

Guckenheimer, J., Oster, G., & Ipaktchi, A. (1977). The dynamics of density dependent population models. *Journal of Mathematical Biology, 4*, 101–147.

Hilborn, R. C. (2002). *Chaos and nonlinear dynamics*. Oxford: Oxford University Press.

Kellert, S. H. (1993). *In the wake of chaos*. Chicago: The University of Chicago Press.

Kellert, S. H. (2008). *Borrowed knowledge: Chaos theory and the challenge of learning across disciplines*. Chicago: The University of Chicago Press.

Li, T.-Y., & Yorke, J. A. (1975). Period three implies chaos. *The American Mathematical Monthly, 82*, 985–992.

Lorenz, E. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences, 20*, 130–141.

Lorenz, E. (1993). *The essence of chaos*. London: UCL Press.

May, R. M. (1974). Biological populations with non-overlapping generations: Stable points, stable cycles, and chaos. *Science, 15*, 645–647.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature, 261*, 459–467.

May, R. M. (1986). The Croonian lecture, 1985: When two and two does not make four: Non-linear phenomena in ecologys. *Proceedings of the Royal Society of London B, 228*, 241–266.

May, R. M., & Osler, G. F. (1976). Bifurcations and dynamic complexity in simple ecological models. *The American Naturalist, 110*, 573–599.

Ruelle, D. (1991). *Chance and chaos*. Princeton: Princeton University Press.

Smale, S. (1967). Differentially dynamical systems. I. diffeomorphisms. *Bulletin of the American Mathematical Society, 73*, 747–816.

Smale, S. (1998). Mathematical problems for the new century. *The Mathematical Intelligencer, 20*, 7–15.

Tucker, W. (2002). A rigorous ODE solver and Smale's 14th problem. *Foundations of Computational Mathematics, 2*, 53–117.

Werndl, C. (2009). Justifying definitions in mathematics: Going beyond Lakatos. *Philosophia Mathematica, 17*, 313–340.

Rudolf Carnap: Philosophy of Science as Engineering Explications

Christopher F. French

1 Introduction

The decision, according to Rudolf Carnap, to adopt a "language structure and, in particular, the decision to use certain types of variables is a practical decision like the choice of an instrument" (1956, 43).[1] What does Carnap mean here by a practical decision? Arguably, he has in mind the kind of decision engineers make on a routine basis; namely, the decision to choose an instrument or tool in order to make something happen. Language structures for Carnap, after all, can be used to formulate theoretical assertions and there is a catalog of language structures to choose from (as evidenced by Carnap's principle of logical tolerance). Accordingly, different language structures can be chosen as a means to formulate theoretical assertions in a way analogous to how an engineer chooses an instrument as a means to make something happen in the world. This analogy is important to Carnap, at least insofar as it helps address worries about the metaphysical commitments of linguistic decisions:

> I admit that the choice of a language suitable for the purposes of physics and mathematics involves problems quite different from those involved in the choice of a suitable motor for a freight airplane; but, in a sense, both are engineering problems, and I fail to see why metaphysics should enter into the first any more than into the second. (Carnap 1956, 43)

[1]I would like to thank Alan Richardson, S. Andrew Inkpen, Taylor Davis and Stefan Lukits for their comments and suggestions on earlier drafts of this paper.

C.F. French (✉)
University of British Columbia, 1866 Main Mall, Vancouver, BC V6T 1Z1, Canada
e-mail: cffrench@interchange.ubc.ca

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_22

For Carnap, certain language structures can be studied and adopted without incurring metaphysical commitments. This is explained by appeal to the engineering analogy. Just as the engineer's choice of an instrument is a practical matter—relative to whether an instrument will get the job done—the decision to use a particular language structure can be investigated without consigning oneself to metaphysical commitments, even if certain metaphysical attitudes or proposals influence one's preference to adopt one structure over others. Rather, each structure in this catalog is to be evaluated as being a better or worse means for achieving one's envisioned theoretical ends.

The purpose of this paper is to clarify the philosophical significance of this engineering analogy. A natural way of trying to do so is the following. Taking Carnap's continuum of inductive methods as an example, the technical modifications and improvements made to that continuum by Carnap and his peers constitute something like an on-going engineering project. Because technical projects are concerned with technical questions, we simply require technical, non-metaphysical answers to these questions. However, I argue that this explanation is unsatisfactory: Carnap's work on inductive logic is analogous to an engineering project not because both are technical projects but because Carnap understands his work in inductive logic as an extension of his work in semantics and the distinction central to the engineering analogy—the distinction between practical and theoretical questions—is made explicitly in his work on semantics.[2] Consequently, in emphasizing the similarities between Carnap's inductive logic and its various extensions and modifications—all the while simultaneously downplaying the dependence of Carnap's inductive logic on semantics—we lose sight of why Carnap thought the practical decisions inherent in the construction of an inductive logic (understood as an explication project) resemble, in a certain sense, engineering problems.

2 Carnap's λ-System

In this section, we quickly review Carnap's work in inductive logic along with several of the related mathematical results. Suppose there are N balls in an urn, each ball is one of finitely many colors, T (≥ 2), and that a sample of S many balls has been observed. Let s_t denote the number of t-colored balls in the sample (where $1 \leq t \leq T$ and $\sum s_t = S$). Inductive methods provide us with rules for attaching numerical values to sentence pairs; for example, a value to the hypothesis that the next ball will be t-colored, H_t, given the total evidence about the sample, E_S. Carnapian confirmation functions provide a way to represent these inductive methods as semantic functions, denoted as $\mathfrak{c}(H_t, E_S)$, defined over the sentences

[2]The claim is not, however, that a theory of semantics is required to make this distinction as Carnap draws a similar distinction between mathematical and empirical geometry (e.g. in Carnap 1939).

H_t and E_S in some object language $\mathfrak{L}$.[3] However, it is far from obvious which of the infinitely many inductive methods $\mathfrak{c}(H_t, E_S)$ should be based on. Fortunately, Carnap's λ-system, Eq. 1 below, captures a particular continuum of inductive methods up to the parameter value λ,[4]

$$\mathfrak{c}_\lambda(H_t, E_S) = \frac{s_t + \lambda/T}{S + \lambda}. \tag{1}$$

The crucial assumptions required for this result are as follows: (i) $\mathfrak{c}$ has to be regular (all non-false sentences have a positive confirmation value), (ii) all sequences of observed balls are exchangeable in the sense it doesn't matter how we index the balls from 1 to N, and (iii) what has come to be known as the sufficiency postulate: for all values of λ, $\mathfrak{c}_\lambda(H_t, E_S)$ is a function of only s_t and S (for details, see Zabell 2005).

These conditions are not entirely historically accurate: they belong to a certain reconstruction of Carnap's λ-system that streamlines the differences between Carnap's inductive logic, understood as an extension of his work in semantics, and a host of results from mathematical probability theory and theoretical statistics. At least in the early 1950s, Carnap's confirmation functions were defined in terms of measure functions and measure functions where defined in terms of the semantic concepts of a state description and the range of a sentence (§18, §§55–56 Carnap 1962).[5] In Carnap (1952), this semantic machinery proved to be slightly problematic. He needed a way to transform confirmation functions into mathematical functions, functions which could then be expressed in terms of a characteristic (or simply, a recursive) function. Carnap's solution to this problem is to make several simplifying assumptions about the enumerative induction situation exemplified by the urn example above. He imposes certain symmetry constraints on the predicates and constants of $\mathfrak{L}$ and the result is the classification of a special class of regular confirmation functions (1952, §§4–5). Carnap then shows that a characteristic function, viz. a function of only the values s_t, T and S (later he learned he could drop T), can be used to parameterize this special class of confirmation functions for any value of λ in $[0, \infty)$ ($\lambda = \infty$ is treated as a special, limiting case) (1952, 16, §§9–10). The result is a continuum of confirmation functions expressed by Eq. 1 above. Consequently, as the right-hand side of Eq. 1 includes only numerical terms, the semantic approach (or what Kuipers 1978 calls the logico-

[3]Here $\mathfrak{L}$ is simply a first-order language with N many named constants representing the balls in the urn and T many basic monadic, "color," properties. Early on, Carnap assumed these properties, expressed as Q-properties, were independent. However, following the work of John G. Kemeny, this earlier assumption was soon relaxed with the technical devices of meaning postulates and families of properties (see Carnap 1971).

[4]For example, Reichenbach's straight rule (which is just the relative frequency s_t/S) is given by Eq. 1 when $\lambda = 0$; but note that $\mathfrak{c}_0$ is not regular.

[5]Later Carnap instead uses models and sigma algebras, see Carnap (1971, §§1, 3).

linguistic approach) turns out to be redundant. The λ-system can be replicated mathematically merely by specifying values for T, S, N and $s_1, \ldots, s_T$.

Seeing as how the semantic gloss can be removed from Carnap's λ-system, we can now view that system from a statistical, Bayesian standpoint. From this point of view, the λ-system isn't new at all—it was already discovered by William Ernest Johnson in the 1920s (see Zabell 2005, 90; Good 1965, 23–30). Moreover, not only are there many ways to generalize the λ-system itself, as detailed in Kuipers (1978), there are several ways to relax Carnap's strict assumptions about $\mathfrak{L}$ to arrive at less idealized inductive systems.[6] Arguably, however, the most significant result is the connection between the λ-system and Bruno de Finetti's representation theorem. Roughly speaking, the conceptual link is that both results assume that sequences of observed samples are exchangeable and the representation theorem tells us how to express the limiting frequency of exchangeable sequences of observations (e.g. infinite sequences of T-many-colored balls from our urn) as a unique mixture of some unknown probability measure. Another way of putting the point is to note that Carnap's λ-system is the finite version of symmetric Dirchlet priors: in the limit, the λ-system corresponds to a probability measure (defined up to a single real-valued parameter) over probability measures which are themselves defined over exchangeable sequences of random variables (see Good 1965; Zabell 2005).

3 Languages as Instruments, Concepts as Tools

As important as these results are for contemporary Bayesian approaches to probability and confirmation theory, they do little to help us understand Carnap's mature philosophical view. Of course, not all of Carnap's peers ignored the possible philosophical implications of his technical projects. Jaakko Hintikka, for example, stresses that his own work in inductive logic is not a mere "tinkering" with Carnap's λ-system but is rather "for all practical purposes a refutation of Carnap's philosophical program in developing his inductivist logic" (Hintikka 1987 302; 305).[7,8] I mention Hintikka's worry for two reasons. First, Hintikka correctly acknowledges a philosophical project underlying Carnap's technical inductive logic project (even if there is disagreement about what that philosophical project is) and second, as Hintikka's use of the "tinkering" locution suggests, Hintikka presumably does not understand himself to be an engineer, or at least not an inductive technician.

[6] For example, Zabell (2005) considers the case where T is infinite (chapter 11) and Skyrms (2012) relaxes Carnap's symmetry assumptions to allow for periodic samples using recurrent Markov chains (see chapters 11 and 12).

[7] Kuipers (1978) recognizes Carnap's project as a philosophical project of concept formation.

[8] Using the semantic concept of a constituent, Hintikka extends the λ-system to two parameters so as to assign non-zero confirmation values to universal generalizations. For details, see Niiniluoto 2011 and the references therein.

Indeed, Carnap understands his work on inductive logic as an explication project, a philosophical project meant to clarify and help systematize the role of the logical concept of probability in inductive reasoning (see French 2015 and Kuipers 2007). In lieu of explaining Carnap's understanding of explication projects in any detail, we instead focus on the central distinction for explication projects: the practical/theoretical distinction introduced in Sect. 1. Since this distinction is well understood within the context of Carnap's work on semantics, we will now briefly turn to his work on semantics in the 1930s and 1940s.

In 1939, Carnap divides the investigation of the language of science into syntax, semantics and pragmatics (see Uebel 2013) and he distinguishes between two kinds of linguistic investigations. The first, descriptive, investigation uses the pragmatic facts of a language already in use to articulate explicit semantic and syntactic rules which seem to be implicit in those facts (1939, 166). The second, conventional, investigation does not concern a language already in use. According to Carnap, here "we are not bound by a previous use of language, but are free to construct in accordance with our wishes and purposes" (166). For example, we could construct a pure semantics with specific "practical applications" in mind, like "making communications or formulating a scientific theory" (166). Importantly, Carnap suggests that both kinds of investigations can be used to study the role that both logic and mathematics have played in "furnish[ing] instruments for deduction, that is, for the transformation of formulations of factual, contingent knowledge" (144). A mathematical or logical theorem applied in scientific contexts is itself "a device or tool" and mathematics "furnish[es] these instruments" for the empirical sciences:

> the mathematician not only produces them for any particular case of application but keeps them in store, so to speak, ready for any need that may arise. (189)

Just as the mathematician furnishes instruments for scientists, Carnap later intimates that the philosopher who engages in the free construction of syntactic or semantical systems may likewise construct logical instruments, instruments which can then be used for certain philosophical ends. Importantly, the construction of a pure system may be "guided," according to Carnap, by the "pragmatical facts" of an historical language (1942, 13). Nevertheless, such systems still do not make theoretical assertions about the world. Pure semantics, says Carnap,

> is not a branch of empirical science; it does not furnish knowledge concerning facts of nature. It is rather to be regarded as a tool, as one among the logical instruments needed for the task of getting and systematizing knowledge. As a hammer helps a man do better and more efficiently what he did before with his unaided hand, so a logical tool helps a man do better and more efficiently what he did before with his unaided brain, that is, by means of instinctive habits rather than through deliberate acts guided by explicit rules. (1943, viii)

This practical/theoretical distinction makes its way into Carnap's talk of explications as follows. There are two active ingredients in Carnap's method of explication, i.e. the method of replacing (if only partially) some inexact or vague concept, the

explicandum, with an exactly defined concept, an *explicatum*.[9] On the one hand, the semantic and syntactic features of the explicatum should be similar to the explicandum; this is a project in descriptive semantics and/or syntax. Nevertheless, the explicandum—as it is employed in everyday conversation via the "unaided brain"—may be an imprecise, clumsy, concept (or cluster of concepts). Thus the explicatum isn't meant to completely mirror the semantic and/or syntactic properties of the explicandum. So on the other hand, we may want to make the explicatum more precise or exact so that particular applications of the explicatum are more efficient, more useful, than what we could accomplish before with the explicandum alone. This is a "conventional" project in pure semantics and/or syntax.

Aside from the first step of clarifying the explicandum, drawing an analogy between pure, or mathematical, geometry and physical geometry, Carnap later explains that explication projects can also be split into a "pure" and "applied" component (see 1971, §4). Using Carnap's work on inductive logic as an example, "pure" inductive logic is the process of defining an inductive logic in a particular logical system. The construction of the λ-system, for example, is a purely logical or mathematical process and the choice of how to define a unique degree of confirmation for every value of λ is a purely practical question. Questions about how to interpret and apply such logics may be informed by the inductive practices and requirements of scientists (see 1953, 195–196).[10] For example, one application of an inductive logic interpreted in terms of betting quotients would be to define concepts of credibility and initial credence functions for use in normative decision theory (e.g. see 1971, §§4,8).[11]

Nevertheless, I argue that it is because explication projects contain this pure, "conventional" ingredient that they resemble engineering projects. Specifically, the question of how to design an explicatum is a practical question. There are always alternative designs to choose from and each design corresponds to a different way of trying to explicate the explicandum. But there is no guarantee that any one explicatum will be more useful in those contexts where the explicandum was ambiguous or vague. For Carnap, no explicatum is "correct": instead, each is only better or worse relative to how (i) similar it is to the explicandum and whether it is (ii) exact, (iii) fruitful and (iv) simple (see 1962, §§1–6). Indeed, Carnap explicitly understands the λ-system itself as a kind of instrument. The question, says Carnap, regarding which value of λ should be used in order to pick an adequate inductive method "is fundamentally not a theoretical question" (1952, 53). Instead, the answer requires a practical decision rather than an assertion:

[9]Importantly, Carnap's views on pragmatics, semantics and syntax precede Carnap's introduction of the explication nomenclature in 1945.

[10]Carnap is explicit in 1962 that his aim is neither to provide confirmations for *entire* scientific theories nor to automate scientific inference, e.g. see his discussion (motivated by results from Alan Turing) of "the impossibility of an automatic inductive procedure" (192 ff.).

[11]Thanks to an anonymous referee for prompting me to clarify this paragraph.

> the adequacy of the choice depends, of course, on many theoretical results concerning the properties of the various inductive methods; and therefore the theoretical results may influence the decision. Nevertheless, the decision itself still remains a practical matter, a matter of *X* making up his mind, like choosing an instrument for a certain kind of work. (1952, 53)

Consequently, an inductive method, says Carnap, is "an instrument for the task of constructing a picture of the world" and this instrument can be changed just as a person "changes a saw or an automobile, and for similar reasons" (55). In fashioning an inductive logic as we would design and construct an instrument, we make our inductive preferences and practices explicit while also keeping in mind how different instruments may more or less satisfy our theoretical aims.

4 Conceptual Engineering

Now that we have some idea of how the practical/theoretical distinction from Carnap's work in semantics comes to play a role in the λ-system, we next discuss the philosophical significance of this analogy. Carnap is well-known for suggesting that philosophical progress can be made in philosophy by transforming traditional philosophical questions into precise, technical, questions. This is Carnap's version of a scientific philosophy: instruments and tools, like logic and mathematics, are borrowed from the sciences and then used to clarify and modify philosophical (and scientific) questions and concepts. The worry, however, is that Carnap is only manufacturing technical answers to technical questions, questions orthogonal to legitimate philosophical problems.

The engineering analogy is used in the recent Carnap reappraisal literature to help explain why Carnap thought his technical projects were still philosophical.[12] Richard Creath, for example, characterizes the engineering analogy as follows:

> philosophers can devise, refine, and explore a variety of conceptual or linguistic frameworks and test their suitability for various practical purposes. These frameworks are tools, so we do not have to prove that they are correct. Nor do we have to agree on which ones to use. We just have to be clear enough to see what follows from what. Then a new result, whether it is a newly clarified concept or a new theorem is a new and permanent and positive addition to our stock of tools. And Carnap can offer the preceding three decades and more in logic as an example of the sort of continuing progress that he is describing. Logicians often disagreed about which systems to use, but they almost never disagreed about what were the results of another's systems. (2009, 211)

[12]Richardson (2013), for example, has suggested that Carnap's treatment of logical languages as instruments was influenced by his earlier interests in metrology. Carus (2007), on the other hand, offers a more systematic, but controversial, account of Carnap as conceptual engineer. A more extensive discussion of the engineering analogy—including how to understand engineering itself—can be found in French (2015).

Linguistic structures, or frameworks, for Carnap, offer a common ground for how philosophers and scientists can communicate their disagreements. Disagreements, for example, about how to define analyticity for logical systems. We gave some idea of what it means for Carnap to treat a linguistic structure as a tool in the last section. In the next section, while keeping in mind Creath's account of the engineering analogy, we will briefly consider an example of Carnap as conceptual engineer.

5 Case Study: Finding Optimal λ-Values

In Carnap (1962), first published in 1950, Carnap suggests that current work on the foundations of theoretical statistics is fractured and piece-meal (513). In short, Carnap recognizes that most statistical frameworks, like Neyman-Pearson hypothesis testing, "are presumably the best instruments for estimating parameter values and testing hypotheses" even though they are fundamentally *not* probabilistic (518). Nevertheless, Carnap suggests his work in inductive logic may bring clarity to this situation: if his inductive logic is adopted, only "one fundamental decision" is required; namely, the decision to choose a confirmation function, a function which can then be used to define other probabilistic and statistical concepts (514).

More specifically, Carnap has in mind the problem of estimation from theoretical statistics (514). Letting $\mathscr{E}$ be an estimator for the parameter θ which represents the empirical phenomenon of interest, suppose our observed data consists in S samples $\mathbf{x} = (x_1, \ldots, x_S)$. The problem of estimation, in a nutshell, is to find estimation functions for which $\mathscr{E}(\mathbf{x})$ is "close" to the actual value of θ, call it $\hat{\theta}$. As Carnap points out, one well known solution to this problem—R. A. Fisher's principle of maximal likelihood—presupposes that $\mathscr{E}$ is unbiased (i.e. the expectation of $\mathscr{E}(\mathbf{x})$ equals $\hat{\theta}$) and as a consequence Fisher's rule presupposes an inductive method; namely, that the probability for each data point x_i is given by Reichenbach's straight rule.

Alternatively, in 1952, Carnap suggests we investigate estimation functions without presupposing any inductive method at all. Specifically, for some fixed state description $\mathbf{k}$, letting $\mathbf{x}$ represent observations about colored balls in an urn and $\hat{\theta}$ the actual frequency of blue balls in $\mathbf{k}$, then for any value of λ consistent with the λ-system, Carnap defines a parameterization of estimation functions as follows,

$$\mathscr{E}_\lambda(\hat{\theta}, \mathbf{k}) =: \sum_{\theta_j} [\theta_j \cdot \mathfrak{c}_\lambda(H_{\theta_j}, E_S)], \tag{2}$$

where E_S is the total evidence of the observed sample of size S and H_{θ_j} is the hypothesis that the actual frequency of blue balls in $\mathbf{k}$ is θ_j. Although the estimation function $\mathscr{E}_{\lambda=0}$ is the only function in Carnap's framework that is unbiased, Carnap goes on to argue that if we measure the adequacy of estimation functions in terms of their mean squared error $\mathfrak{w}^2$ (where the error of $\mathscr{E}_\lambda$ is the difference between

$\mathscr{E}_\lambda(\mathbf{x})$ and $\hat{\theta}$) then there exists—for fixed sample size S and some arbitrarily large, "non-extreme," state description $\mathbf{k}'$—an "optimal" value of λ, call it λ^*, such that $\mathfrak{w}^2[\mathscr{E}_{\lambda=\lambda^*}(\hat{\theta}, \mathbf{k}')] < \mathfrak{w}^2[\mathscr{E}_{\lambda=0}(\hat{\theta}, \mathbf{k}')]$.[13] In other words, there is some "universe" characterized by $\mathbf{k}'$ for which some biased estimate is "closer" to the actual value of $\hat{\theta}$ than the unbiased estimator, $\mathscr{E}_{\lambda=0}$ (see 1952 §§22–24).

The philosophical relevance of Carnap's result is that he describes it as a "neutral" investigation: unlike Fisher's maximal likelihood method, Carnap's method presupposes that no inductive method, like the straight rule, is correct (1952, 60). Indeed, unlike the Bayesian variants of Carnap's λ-system, Carnap's parameterization of inductive methods itself is an instrument, an inductive technology, which has been constructed from a language framework. Consequently, as we saw in Sect. 3, the choice of this framework over others is a practical matter and as such, according to Carnap, we incur no ontological commitment when we investigate whether certain estimators are more or less successful within a range of mathematically constructed state descriptions, each of which correspond to a different "universe." Which state description we adopt is a logical matter and after that it is just a question of figuring out how to use a bit of differential calculus to show which values of λ are optimal. Moreover, it is precisely because this is the case that Carnap says that we "must necessarily refrain from making any judgment concerning the success of an inductive method in the total actual world" (60).

Questions about which estimation functions we should actually use in real-life cases concern any number of methodological, non-logical and empirical questions about how to apply our inductive logic (see Carnap 1962, §44; see Uebel 2013 for a discussion about Carnap's intellectual division of labor between logicians, scientists and philosophers). Of course, there will be many different ways of deciding how to apply this inductive instrument to real-world problems; nevertheless, however we do so, Carnap's point is that there is no deep metaphysical mystery for how this could be possible. Indeed, it is not essentially different from the case of applying mathematical geometry to empirically measure physical objects.[14]

[13] Technicalities aside, the main idea is that each state description $\mathbf{k}$ has a certain degree of order, which Carnap understands as an explication of the regularity of the actual world. Letting $r_1, \ldots, r_T$ be the actual relative frequencies of the T many monadic properties in $\mathbf{k}$, Carnap explicates the degree of order of $\mathbf{k}$ as the squared sum of the r_i's, $\sum r_i^2$. The value $\hat{\lambda}$ is optimal for $\mathbf{k}$ when $\hat{\lambda} = (1 - \sum r_i^2)/(\sum r_i^2 - 1/T)$. Moreover, $\mathbf{k}$ is "non-extreme" in the sense above means that $\sum r_i^2 \neq 1$; also, note that $\hat{\theta}$ is equal to one of the r_i's (see §24). Carnap had plans to published more on this concept but did not; although see the manuscripts regarding this concept and the related abstract concept of entropy, RC 086-07-01 and RC 080-15-01, at the Rudolf Carnap archives at the Archives of Scientific Philosophy (ASP) at the University of Pittsburgh.

[14] Although Carnap does have written correspondence with statisticians like L. J. Savage, de Finetti, Jerzy Neyman and Harold Jeffreys, besides Savage pointing out to Carnap in 1952 that λ cannot vary with sample size (ASP RC 084-52-25), there seems to have been little interest among statisticians in Carnap's λ-system (but see Good 1965). In this sense, Carnap's intervention in the problem of estimation is a failure of sorts.

6 Conclusion

We have just seen an example of how Carnap suggested his work on inductive logic could possibly provide a "neutral" framework to help clarify the problem of estimation in theoretical statistics. I suggest that this "neutral" framework is best understood as a kind of conceptual engineering. The upshot is that Carnap's technical work on estimation functions is more than a mere technical project. Instead, the framework provides a concrete example for what philosophers, according to Carnap, should attempt to achieve: namely, to use technical machinery to help clarify and systematize ambiguous concepts, especially foundational concepts central to both traditional philosophy and science.

However, I do not seek to downplay the significance of any of the extensions or modifications made to Carnap's work on inductive logic discussed in Sect. 2. Carnap's inductive logic, although it provided the foil for important philosophical projects like Hintikka's work on inductive logic, proved practically difficult to combine with ever-increasingly sophisticated results in Bayesian statistics and it makes sense why one would want to use these more recent results to streamline Carnap's inductive logic. Of course, Carnap would have expected nothing less: he was, after all, not one to set up prohibitions against the use of technical machinery to help address philosophical problems.

Nevertheless, just because Carnap failed to convince his philosophical and scientific peers about the usefulness of his explications that does not mean that the engineering standpoint is flawed. As a way of mediating disagreements and making progress in philosophy, a conceptual engineer crafts concepts with the aim of bringing clarity to our intellectual lives by designing and constructing the concepts we use to make theoretical assertions. As engineering problems, our failures to address philosophical worries become teaching moments: we can learn from our mistakes and try to design concepts more amenable to the sensibilities of our philosophical peers. Crucially, the upshot of Sect. 5 is that we can accomplish all this without the requirement that we must first pinpoint and justify our perceived ontological commitments. Indeed, Carnapian conceptual engineering provides an answer to the question whether technical projects should have a place in philosophy at all. Just as the λ-system is a means by which a philosopher can help adjudicate inductive disputes, in adopting something like Carnap's practical/theoretical distinction for our own preferred technical machinery, we may come to have a clearer picture about how to design our own concepts in order to clarify, or even ameliorate, our own philosophical disputes.[15]

[15] A much less condensed discussion of the ideas in this paragraph can be found in French (2015). In future work I plan to extend this view of Carnap as conceptual engineer to various other ares of contemporary philosophy of science, for example, with various notions of "progress" in empirical concept formation (e.g. see Kuipers 2007); with other kinds of engineering relevant for philosophy, viz. in terms of Herbert Simon's notions of "bounded rationality" and "satsificing" (e.g.

References

Carnap, R. (1939). *Foundations of logic and mathematics*. Chicago: University of Chicago Press.
Carnap, R. (1942). *Introduction to semantics*. Cambridge: Harvard University Press.
Carnap, R. (1943). *Formalization of logic*. Cambridge: Harvard University Press.
Carnap, R. (1952). *The continuum of inductive methods*. Chicago: University of Chicago Press.
Carnap, R. (1953). Inductive logic and science. *Proceedings of the American Academy of Arts and Sciences, 80*(3), 189–197.
Carnap, R. (1956). *Meaning and necessity: A study in semantics and modal logic* (2nd ed.). Chicago: University of Chicago Press.
Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago: University of Chicago Press.
Carnap, R. (1971). A basic system of inductive logic, part I. In J. Richard & R. Carnap (Eds.), *Studies in inductive logic and probability* (Vol. 1, pp. 34–164). Los Angeles: University of California Press.
Carus, A. W. (2007). *Carnap and twentieth-century thought*. Cambridge: Cambridge University Press.
Creath, R. (2009). The gentle strength of tolerance: The logical syntax of language and Carnap's philosophical programme. In P. Wagner (Ed.), *Carnap's logical syntax of language* (pp. 203–214). Basingstoke: Palgrave-MacMillan.
French, C. F. (2015). *Philosophy as conceptual engineering: Inductive logic in Rudolf Carnap's scientific philosophy*. PhD thesis, University of British Columbia.
Good, I. J. (1965). *The estimation of probabilities: An essay on modern Bayesian methods*. Cambridge: MIT.
Hintikka, J. (1987). Replies and comments. In R. J. Bogdan (Ed.), *Jaakko Hintikka* (pp. 277–344). Dordrecht: D. Reidel.
Kuipers, T. A. F. (1978). *Studies in inductive probability and rational expectation*. Dordrecht: D. Reidel.
Kuipers, T. A. F. (2007). Explication in philosophy of science. In T. A. F. Kuipers (Ed.), *Handbook of the philosophy of science. General philosophy of science – Focal issues* (pp. vii–xiii). Amsterdam: Elsevier.
Morgan, M. S. (2012). *The world in the model: How economists work and think*. Cambridge: Cambridge University Press.
Niiniluoto, I. (2011). The development of the Hintikka program. In D. M. Gabby, S. Hartmann, & J. Woods (Eds.), *Handbook of the history of logic: Inductive logic* (Vol. 10, pp. 311–356). Oxford: North Holland.
Richardson, A. W. (2013). Taking the measure of Carnap's philosophical engineering: Metalogic as metrology. In E. H. Reck (Ed.), *The historical turn in analytic philosophy* (pp. 60–77). Basingstoke: Palgrave-Macmillan.
Skyrms, B. (2012). *From Zeno to Arbitrage: Essays on quantity, coherence, and induction*. Oxford: Oxford University Press.
Uebel, T. (2013). Pragmatics in Carnap and Morris and the bipartite metatheory conception. *Erkenntnis, 78*(3), 523–546.
Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge: Harvard University Press.
Zabell, S. (2005). *Symmetry and its discontents*. Cambridge: Cambridge University Press.

see Wimsatt 2007); or finally with how social scientists, especially economists, use and employ mathematical models (e.g. see Morgan 2012).

Robustness, Diversity of Evidence, and Probabilistic Independence

Jonah N. Schupbach

1 Robustness Analysis in Science

To verify that results are not simply artifacts of the particular means used to detect them, scientists often attempt to duplicate those results using other, diverse means. To the extent that a result is detected via numerous, diverse means, it is said to be *robust*. *Robustness analysis* (henceforth, "RA") is a mode of reasoning in which one supports a hypothesis via an analysis of the conditions under which a result is robust.

Examples of RA from scientific practice abound. Famously, biologist Richard Levins (1966) proposes RA as a general means for deciphering, when using simplified models to study complex systems, whether a result "depends on the essentials of a model or on the details of the simplifying assumptions:"

> [W]e attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies.

As a specific example of this use of RA in modeling, Weisberg and Reisman (2008) discuss work in support of the Volterra Principle:

Volterra Principle *Ceteris paribus*, if a two-species, predator-prey system is negatively coupled, then a general biocide will increase the abundance of the prey and decrease the abundance of predators.

The earliest predator-prey models demonstrating the Volterra Principle included several unrealistic, simplifying assumptions – including single growth- and death-

J.N. Schupbach (✉)
Department of Philosophy, University of Utah, Salt Lake City, UT 84112, USA
e-mail: jonah.n.schupbach@utah.edu

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_23

rates for prey and predators respectively as well as linear functions relating prey capture-rates to number of predators and predator birth-rates to number of captures. As Weisberg and Reisman emphasize, however, the Volterra Principle can be demonstrated robustly across more complicated models that do away with various of the simplifying assumptions. Some of these models, for example, "add in terms representing predator satiation, the ability of prey to seek cover, multiple sources of food for the predator, or even complex adaptive behaviors such as learning" (116).

In this example, the robust result is the observed qualitative behavior of various models – commonly interpreted as the increased relative size of prey population when a general biocide is introduced. The diverse means which detect this result are the mathematical models themselves. And the hypothesis most apparently supported by the fact that this result is robust is the Volterra Principle – the biological claim, not to be confused with claims about the mathematical representations of biological systems.

The case of Brownian motion provides a rather different example of RA in science – also commonly discussed by philosophers of science. When suspended in a fluid medium, sufficiently small particles display continual and seemingly random movements. Upon discovering this phenomenon, the botanist Robert Brown surmised that the motions were due to the particular (uniquely shaped) pollen granules he was observing. However, he found that the movements persisted across experiments using other particles – first using other types of pollen, then other organic materials, and eventually using inorganic particles. Over the next 75 years, other experimenters showed that the "Brownian motion" was also robust over changes in the fluid medium, container used, means of suspending the particles, environmental conditions around the container, and so on. Eventually, scientists working in the wake of Einstein's *annus mirabilis* appealed to the robustness of the Brownian motion in order to support the idea that there were deeper, unobservable agitations of molecules within the medium – just as the evident rocking of a far off ship betrays imperceptibly distant waves on the sea (Perrin 1913, 83).

In this case, the Brownian motion is the result detected robustly. The diverse means of detection are the various experiments used; regarding these experiments, the Brownian motion is notably robust across certain changes to the experimental apparatus (type of particle, medium, container, lighting, etc.) and sensitive to others (size of particle, temperature of medium). And it is upon analyzing these conditions of robustness that Perrin (1913, 86) says we are "forced to conclude," consonant with Einstein's molecular explanation, that there are internal, unobservable movements in the medium.

Many other example RAs include cases from cognitive psychology (Crupi et al. 2008; Stolarz-Fantino et al. 2003), "arguments from coincidence" in physics (Cartwright 1991; Hacking 1983; Mayo 1996), experimental biology (Culp 1994), climate science (Lloyd 2010; Parker 2011), and modeling in economics (Kuorikoski et al. 2010; Woodward 2006). As should be evident from the range of these cases, I mean for the terms "results" and "means of detection" to be quite generic. The results in question could be observations, measurements, predictions, theorems,

and so on. Correspondingly, the means of detecting such results could include experiments, laboratory instruments, sensory modalities, derivations (from axioms, models, theories, etc.), axiomatic systems, computer simulations, and formal models amongst other things.

2 Evidential Diversity and RA-Diversity

The intuition motivating the use of RA is that we can gain confirmation through diversity; certain hypotheses (e.g., Einstein's molecular explanation of Brownian motion) are supported to the extent that a result proves robust, and results are robust to the extent that we detect them in diverse ways. But what precise sense of diversity is involved in RAs? Philosophers have offered many distinct accounts of evidential diversity. And many of these accounts plausibly capture legitimate senses in which we speak of evidence as being diverse. But is there a single sense of evidential diversity that drives our reasoning in RAs? For the sake of this paper, we will work on the optimistic assumption that there is.

The hope is that a precise account of such "RA-diversity" would illuminate the normative import of actual RAs. But in order to stand any chance of doing so, such an account must be held accountable to scientific practice. A particular account of evidential diversity may, for example, specify precise conditions under which diverse bodies of evidence provide strong confirmation for relevant hypotheses. But such an account will clearly not illuminate RA if it relies on a notion of diversity that does not fit with actual cases of RA in science. Such an account may shed light on the confirmational import of certain diverse bodies of evidence, just not on RA-diverse bodies of evidence.

This main section of the present paper applies this consideration in criticizing the most common formal accounts of RA-diversity. These formalize diversity using probabilistically-precise notions of independence. Moreover, several of these accounts imply interesting senses in which diverse bodies of evidence may be specially confirmatory. The problem is that these accounts fail to capture paradigmatic cases of RA from science. To show this in each case, I will return to the above examples of Brownian motion and the Volterra Principle.

2.1 Unconditional Probabilistic Independence

As a first attempt at explicating RA-diversity, we might take a cue from Levins's quote and surmise that some precise notion of "independence" is at work. Most simply, one might say that if two means of detection are RA-diverse in the relevant sense, then they are (unconditionally) probabilistically independent of one another.

To make this idea more precise, let R be a proposition describing the result that has been robustly detected by various means. Then, let us denote the proposition that this result is detected using the k'th means of detection as R_k. According to the *unconditional probabilistic independence* account, if two means of detection are RA-diverse, then the fact that R is detected via means i should have no bearing whatever on the probability that R will be detected using means j: $Pr(R_i \& R_j) = Pr(R_i) \times Pr(R_j)$ – which (assuming that $Pr(R_i), Pr(R_j) > 0$) entails that $Pr(R_i) = Pr(R_i|R_j)$ and $Pr(R_j) = Pr(R_j|R_i)$.[1]

In their critique of Levins's discussion of RA, Orzack and Sober (1993, 539–40) consider and quickly dismiss this explication; they argue that, by requiring the various models to share "a common biological assumption," Levins's " 'Protocol' for the discovery of robust predictions guarantees that the models under consideration are *not* independent."[2] In a bit more detail and put more explicitly in Bayesian terms, when such a model implies a particular result in RA settings, we consider it possible (and perhaps even plausible) that the result is driven by the essential core of the model – i.e., Levins's common biological assumption. But then whether or not we get a result from one of the models will manifestly provide relevant information with regards to whether we will get the result from another model with the same common core. Typically, the fact that we have detected R with one such model will increase the probability that we will detect R using another: $Pr(R_i) < Pr(R_i|R_j)$. Importantly, this may be true despite the fact that these models are considered diverse for the sake of RA.

For similar reasons, RA-diverse experiments in the case of Brownian motion also fail to be unconditionally independent. Take any two of these experiments, say those suspending dust particles in water and those suspending them in ethanol. Although these are diverse in the relevant sense that makes it appropriate for scientists, like Perrin, to cite them as part of their RA, the respective results of these experiments may strongly inform one another. This will be the case, in fact, so long as one allows that other factors (besides whether to use water or ethanol as the medium) may potentially influence whether one observes the result. These experiments actually share in common the vast majority of their respective traits – type of particle used, means of suspending the particle, lighting conditions, etc. – which may all be seen as potentially relevant in affecting the result. But then, observations of Brownian movements in water may greatly increase the probability that one will observe Brownian movements in ethanol. In this case again, perfectly RA-diverse means of detection may be such that $Pr(R_i) < Pr(R_i|R_j)$.

[1] For clarity and ease of exposition, I leave the background beliefs term implicit in all Bayesian formulations.

[2] Orzack and Sober also criticize an alternative explication according to which two models are diverse only if they are *logically* independent. The fact that RA-diverse models may involve contrary simplifying assumptions spells trouble for this account; e.g., "A model with the assumption of random mating is not logically independent of a model with the assumption that mating is assortative; the reason is that the truth of one entails the falsity of the other."

2.2 Reliability Independence

While this initial effort thus fails, there are subtler ways one can attempt to use probabilistic independence to explicate RA-diversity. Wimsatt (1994, 197) offers such an account, proposing "that the *probability of failure* of the different means of access should be independent." This account arguably doubles as a more accurate interpretation of Levins's thought that RA requires "independent *lies*." The lies, the ways that each means of detection could lead us astray, are the things that are required to be independent between RA-diverse means of detection.[3]

This *reliability independence* account is importantly distinct from the unconditional independence account above. Instead of enforcing the stringent condition that the results of the various means of detection be entirely irrelevant to one another, this account just requires that if the means in question lead us astray, they do so for independent reasons. More precisely, learning that one of our means of detection has misled us has no effect on the probability that the other means of detection will mislead us. Each means of detection is or isn't reliable, independent of the others.

One nice feature of this account is the straightforward way in which it reveals the epistemic appeal of diversity. The justification that a hypothesis receives from evidence that is diverse in this sense has all the logical advantage of webs over chains. Whereas a linear chain of justification can be no stronger than its weakest link, a web of independent lines of justification is no weaker than its strongest member. Wimsatt (1981, 49–50) offers a quick probabilistic demonstration of this as follows: Assume that we have n means, all of which detect a result. Now assume that these means are reliability independent. Naturally, these means are imperfect, and so each may lead us astray with some probability; for simplicity, assume that they each may lead us astray with the same probability p_0. Now, if the common result these means are all detecting is misleading, then all n means of detection are going astray. Because they do so independently of one another, we know that the probability of this happening is $p_p = p_0^n$. Wimsatt concludes, "But p_0 is presumably always less than 1; thus, for $n > 1$, p_p is always less than p_0. Adding alternatives for redundancy always increases reliability."

Unfortunately, while reliability independence manifestly explicates an important notion of evidential diversity, it too fails to capture the notion of RA-diversity. First, return to the example of Brownian motion, and again consider any two of the RA-diverse means of detection used by Brown; this time, let us compare experiments in which a variety of pollen granules were suspended in water with those in which a variety of inorganic materials were suspended in water. These experiments are again cited as diverse in the sense required for RA. Yet, their respective reliabilities surely have a bearing on one another. To be sure, they could be unreliable for different reasons. But there are any number of *common* reasons that they might

[3]Bovens and Hartmann (2003, 96–97) offer an in-depth formal exploration of this notion of evidential diversity, and Kuorikoski et al. (2010, 544–45) follow Wimsatt in adopting this account as an explication of RA-diversity.

be unreliable too; there are many possible confounding factors that could be driving the result in both cases. Both could be misleading us due to the way the particles are being suspended, due to the use of the same medium, due to the use of the same environmental conditions surrounding the apparatus, etc. To the extent that we are aware of such overlapping sources of potential unreliability, learning that one of these experiments is leading us astray provides relevant information when deciding whether to trust the other. In particular, such information will often greatly reduce our estimate of how reliable the other is.

The reliability independence account encounters the same problem when trying to model RA-diversity in examples from modeling. Two RA-diverse predator-prey models that demonstrate the Volterra Principle may differ only on whether they involve a particular simplifying assumption, say the assumption that prey capture-rate increases linearly with number of predators. A model that is more realistic in this one regard and the fully simplified model will share many potential sources of unreliability when it comes to modeling the complex predator-prey dynamics (e.g., not allowing prey to take cover or learn). But then learning that one of the models is unreliable will potentially greatly increase our confidence that the other is too. In general, fully RA-diverse means of detection can nonetheless be susceptible to many of the same potential confounds; in such cases, learning that one of our means of detection is unreliable will often greatly increase the likeliness that other of our means of detection is similarly unreliable.

2.3 Confirmational and Conditional Independence

Lloyd (2009, 2010) has recently proposed a third independence-based account worth considering here. She proposes that RA-diversity amounts to *confirmational independence*, as explicated by Fitelson (2001). This sense is defined relative to a particular hypothesis (call it H), which we may think of as the hypothesis intuitively supported via the RA. Two means of detection are RA-diverse, according to this account, only if their results incrementally confirm/disconfirm H (raise/lower H's probability) to the same extent regardless of whether we have detected and learned the results using the other means. More formally (using the notation we introduced in Sect. 2.1 above, and where c stands in for an adopted Bayesian measure of incremental confirmation): if the ith and jth means of detection are RA-diverse with respect to H, then $c(H, R_i|R_j) = c(H, R_i)$ and $c(H, R_i|R_j) = c(H, R_j)$.[4]

As with Wimsatt's account of evidential diversity, this idea nicely illuminates the normative appeal of diversifying our evidence. Accepting any of the most defensible and popular Bayesian measures of confirmation as c, and assuming that each detection of R individually confirms H to some extent, one can prove that

[4]Notation: $c(x, y)$ measures the degree of confirmation that y lends to x; $c(x, y|z)$ measures the degree of confirmation that y lends to x, conditional on (or given that) z.

confirmationally independent means of detection jointly confirm H to a greater extent than either means of detection does individually: $c(H, R_i \& R_j) > c(H, R_i)$ and $c(H, R_i \& R_j) > c(H, R_j)$ (Fitelson 2001, S131).

Before evaluating this account, it is worth mentioning that confirmational independence has a direct connection to conditional probabilistic independence, relative to H. As Fitelson (2001, S129) clarifies, "screening-off by H of R_i from R_j is a sufficient condition for R_i and R_j to be mutually confirmationally independent regarding H."[5] By "screening-off," Fitelson has in mind the standard Reichenbach (1956, 158–59) notion, implying the dual conditional independencies: $Pr(R_i \& R_j|H) = Pr(R_i|H) \times Pr(R_j|H)$ and $Pr(R_i \& R_j|\neg H) = Pr(R_i|\neg H) \times Pr(R_j|\neg H)$.

Unfortunately, confirmational independence also does not fit with the notion of RA-diversity. Consider again two experiments from the Brownian motion case. Let R_1 describe the fact that we have observed Brownian motion using the uniquely shaped and sized pollen granules of *Clarkia pulchella* (the wildflower first used by Brown in his experiments), and let R_2 be the proposition that we have witnessed the same motions using other types of pollen. While these two results are diverse in the sense that makes them crucial to establishing the robustness of Brownian motion (and both mentioned explicitly as such by Perrin), they are evidently not confirmationally independent regarding H: Perrin's inferred hypothesis that there are unobservable movements internal to fluid media. To assert that they are would be to claim that his hypothesis is supported to the same extent by R_1, regardless of whether we know R_2. But while H may be strongly supported by experiments observing the jostling of granules of a particular type of pollen, it plausibly does not gain nearly so much support from such an observation if one has already witnessed the jostling using several other types of pollen: $c(H, R_1|R_2) < c(H, R_1)$. On the contrary, the more pollens that we have already observed in motion, the less a confirmatory impact on H future experiments using pollens will have.

The following observation helps us to see why this account does not work from another angle. The fact that these diverse means of detection are not confirmationally independent regarding H implies that their results also will not be screened-off by H. Here, we can pinpoint the feature of screening-off that generally will not be satisfied by these experiments, the clause that asserts that R_1 and R_2 should be independent conditional on $\neg H$: $Pr(R_1 \& R_2|\neg H) = Pr(R_1|\neg H) \times Pr(R_2|\neg H)$. If H is false, there remain many potential reasons why we might see particles dance about in fluids. Take for example the idea H' that this motion is due to the nature of the suspended particle. Conditional on H', the observation of Brownian motion using various pollens will greatly increase the probability of witnessing it in other pollens: $Pr(R_1|H') \ll Pr(R_1|H' \& R_2)$. After all, on this hypothesis, this motion is attributable to some aspect of the suspended particle; but then witnessing it across samples of pollen will make us more confident that all pollens

[5] I have replaced Fitelson's notation with our own. It should be noted that Fitelson suggests this relation as a condition of adequacy on measures of confirmation, as opposed to proving and presenting it as a theorem that follows robustly (!) for all candidate measures.

share the relevant attribute (e.g., the sexual drive or vital force inherent in the particles). More generally, given that H is false, we might still observe R according to several alternative possibilities. And RA-diverse means of detecting R can be probabilistically relevant to one another conditional on these other possibilities.

Similar points weigh against the idea that confirmational or conditional independence explicates RA-diversity in cases from modeling. For example, conditional on the Volterra Principle being false, there could be several reasons why our models are displaying qualitative behavior interpreted in accordance with this principle. For example, perhaps this behavior is in part an artifact of the unrealistic assumption that prey are borne at a single constant rate. Conditional on the hypothesis that this partially drives our result, two RA-diverse models that both assume single growth-rates for prey (e.g., two models differing only on whether they represent predator satiation) may be substantially probabilistically relevant to one another; if one provides the result, this may greatly increase the probability that the other will too.

2.4 Partial Independence?

One might think that the problem is just that we have framed the above accounts as requiring *full* unconditional, reliability, or confirmational independence. But perhaps we can make these accounts more defensible by adjusting them to measure degrees of RA-diversity. Two means of detection are RA-diverse, we might say, *to the extent* that they approach full unconditional, reliability, or confirmational independence. Wimsatt (1981, 46) may have just this sort of move in mind when he writes, "All these procedures require at least *partial independence* of the various processes across which invariance is shown."

But note that the reasons above for why these accounts fail have little to do with the fact that we require full independence. The problem, in other words, is not that the RA-diverse means of detection in these paradigmatic cases fall just short of full independence in one of the three senses. In fact, we have seen that means of detection that are recognizably and clearly RA-diverse may not even come remotely close to being independent in any of the above three senses. Nor is it at all clear that we would end up with means of detection that are more RA-diverse if we sought those that came closer to full unconditional, reliability, or confirmational independence. In fact, in all of the examples proffered, the means of detection are intuitively fully diverse in the sense required for them to do their work in RA. When Perrin cites experiments detecting Brownian motion using organic particles, and then those using inorganic particles, there is a sense in which these means of detection are perfectly diverse in the sense needed for these to have their respective roles in Perrin's larger RA. And there is no clear reason to think that Perrin's cited means could have been improved in their RA-diversity roles had they been less dependent in one of the above probabilistically precise senses.

What is this general role that means of detection are meant to perform by virtue of their RA-diversity? The answer that I want to explore is, in a word, *elimination*. While the experiments that Perrin cites and the models used to demonstrate the Volterra Principle are actually, in several cases, overall quite similar to one another, they inevitably remain distinct in ways that make them useful for ruling out H's potential competitors. In fact, there is already an account of evidential diversity that fits well with this eliminative idea. The following, closing section briefly explores this account and whether it holds more promise as an explication of RA-diversity.

3 Toward an Alternative Explication of RA-Diversity

Horwich (1982, 118–22) proposes an account of evidential diversity. While this account is probabilistic, it does not make use of probabilistic independence. The central notion in Horwich's account is instead that of *elimination*; diverse bodies of evidence, according to Horwich, "tend to eliminate from consideration many of the initially most plausible, competing hypotheses" (118). Probabilistically, Horwich represents "initially plausible" competing hypotheses as those with substantial prior probabilities, and he identifies diverse evidence with evidence that takes low likelihoods conditional on competing hypotheses.

Horwich argues for the normative appeal of diverse evidence using his account in the following way. Let E_D describe a more *eliminatively diverse* body of evidence than E_N relative to our favored target hypothesis H_1 and its competitors $H_2, H_3, \ldots, H_k$. We can compare the probabilistic effects of both sets of evidence on H_1 by comparing $Pr(H_1|E_D)$ to $Pr(H_1|E_N)$. Horwich stipulates that the alternative hypotheses form a partition $\{H_1, H_2, \ldots, H_k\}$ and that H_1 implies E_N and E_D, so that $Pr(E_N|H_1) = Pr(E_D|H_1) = 1$. Under these conditions, we have:

$$\frac{Pr(H_1|E_D)}{Pr(H_1|E_N)} = \frac{Pr(H_1)}{Pr(H_1)} \times \frac{Pr(E_D|H_1)}{Pr(E_N|H_1)} \times \frac{Pr(E_N)}{Pr(E_D)} = \frac{Pr(E_N)}{Pr(E_D)}$$

$$= \frac{Pr(H_1) + Pr(H_2)Pr(E_N|H_2) + \ldots + Pr(H_k)Pr(E_N|H_k)}{Pr(H_1) + Pr(H_2)Pr(E_D|H_2) + \ldots + Pr(H_k)Pr(E_D|H_k)}.$$

Comparing like terms between the numerator and denominator of this ratio, the only terms that may affect a difference between $Pr(H_1|E_D)$ and $Pr(H_1|E_N)$ are the likelihoods relative to H_1's competitors. Consequently, to the extent that all of those hypotheses with considerable values of $Pr(H_i)$ are such that $Pr(E_D|H_i) < Pr(E_N|H_i)$, it will tend to be the case that $Pr(H_1|E_D) > Pr(H_1|E_N)$. But that is just to say that the more eliminatively diverse the evidence in this case, the more confirmation it will tend to bestow upon H_1.

If we use this account to explicate RA-diversity, we have that means of detection (cited in a RA to H) are diverse insofar as they are able to rule out H's competitors.

On such an account, it is not so important that means of detection are *strongly* diverse or sufficiently distinct in some sense separated from considered hypotheses. What really matters for RA-diversity is that the means (which may actually be quite similar in most respects) are different in just the sense required to rule out *H*'s salient competitors. Accordingly, when seeking to increase the RA-diversity of our evidence, we search for a new way of detecting *R* that rules out some of *H*'s still-standing competitors.

Such an eliminative account of RA-diversity makes better sense of standard cases of RA in science. Many of the RA-diverse means of detecting Brownian motion are overall just not that diverse; indeed, these means may be identical in all respects other than some modest change – e.g., in the particle suspended or mode of suspending it. This is why accounts that require RA-diverse means to be strongly diverse (often in a sense that pays no attention to the relevant hypotheses) run quickly into trouble in this case. These means of detection are clearly eliminatively diverse, however. When Perrin cites experiments on *Clarkia pulchella*, and then increments the RA by citing experiments on other varieties of pollen, he is not doing so because these experiments are strongly heterogeneous overall, but because they are *relevantly* different than one another. The latter rules out a potential confounding hypothesis left standing by the first – viz., that the motion is attributable to the unique form of *Clarkia pulchella* granules.

Similarly, when seeking to confirm the Volterra Principle, RA-diverse models may be identical but for some modest difference. By utilizing these RA-diverse models, we rule out confounding hypotheses pertaining to our result left standing by any subset of the models used alone. Notably, we alleviate worries that our result is an artifact of a simplifying assumption common to some subset of our models by duplicating that result using a new model that does not share that assumption.

Though the eliminative account provides a prima facie more promising approach to explicating RA-diversity, much work remains for any satisfactory account in this vein. Perhaps most obviously, we would ultimately like a more subtle demonstration of the normative impact of eliminative diversity as it relates to RAs. There are several features of Horwich's demonstration that make it less informative regarding RAs. First, the general setting of Horwich's result has us comparing two bodies of evidence, one more and one less diverse, at the end of the collection process. But in practice, we are rarely at the end of an RA, and we are not directly interested in comparing two hypothetical bodies of evidence. The normative intuition that we would like to test is that *H* receives more confirmation with each increment of RA; consequently, a more informative account would examine the confirmatory effects on *H* of adding RA-diverse means of detecting the result to our working body of evidence.

Second, Horwich's demonstration hinges on some very specific problematic assumptions. In RAs, it is not obvious that we should require *H* to imply the detected results in question; this assumption will either need to be weakened or it will require further motivation. Nor is it obvious that all RAs involve hypotheses that compete in the sense of being mutually exclusive; in fact, in the Brownian motion case, Perrin's favored hypothesis is consistent with many of the competing hypotheses that get

ruled out (e.g., it's possible that the motion is affected both by unobservable motions internal to the medium and by a vital force inherent in the particle used). Horwich assumes too that the hypotheses before us exhaust the possibility space (they form a partition). But of course, in actual cases of RA, more often than not, this is not the case. And we would accordingly like an account that informs us of the epistemic import of RA-diversity in cases involving a catch-all hypothesis.

Third, Horwich's account tells us that, under all of the above conditions, more diverse evidence does indeed tend to bestow more confirmation on the relevant sort of hypothesis. But it would be nice if our account could tell us more than this. Is RA confirmatory under other conditions? And what determines the extent of confirmation that an increment of RA provides?

Finally, Fitelson (1996) suggests that Horwich's account is more properly viewed as an explication of the logical effects of diverse evidence than an explication of diverse evidence. There is an "intuitive notion" of evidential diversity that underlies and motivates Horwich's discussion of eliminative diversity. However, argues Fitelson, this intuitive notion proves elusive, and so Horwich's account is at best incomplete. A more appealing account of RA-diversity might provide the missing pieces here, starting from a more intuitive notion of diversity and then showing that something like Horwich's eliminative account captures in part the logical implications of such diversity. All of these considerations point to ways in which a fuller account of RA-diversity along the eliminativist lines will need to expand upon Horwich's account.[6]

Acknowledgements I am grateful for the helpful conversations I have shared on this topic with Aki Lehtinen, Chiara Lisciandra, Gerhard Schurz, Jacob Stegenga, and Ioannis Votsis. Also, thanks to two anonymous referees for their helpful suggestions, which allowed me to improve an earlier draft of this paper. Research for this article was supported by an Aldrich Fellowship from the University of Utah's Tanner Humanities Center, and was conducted during a visit to the Düsseldorf Center for Logic and Philosophy of Science.

References

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

Cartwright, N. (1991). Replicability, reproducibility, and robustness: Comments on Harry Collins. *History of Political Economy, 23*(1), 143–155.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking and Reasoning, 14*(2), 182–199.

Culp, S. (1994). Defending robustness: The bacterial mesosome as a test case. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1994*, 46–57. Contributed papers.

Fitelson, B. (1996). Wayne, Horwich, and evidential diversity. *Philosophy of Science, 63*(4), 652–660.

[6] I have since developed such an account in (Schupbach Forthcoming).

Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science, 68*(3), S123–S140.
Hacking, I. (1983). *Representing and intervening*. Cambridge: Cambridge University Press.
Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *British Journal for the Philosophy of Science, 61*(3), 541–567.
Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421–431.
Lloyd, E. A. (2009). Varieties of support and confirmation of climate models. *Proceedings of the Aristotelian Society, Supplementary Volumes, 83*, 213–232.
Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science, 77*(5), 971–984.
Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.
Orzack, S. H., & Sober, E. (1993). A critical assessment of Levins's *The Strategy of Model Building in Population Biology* (1966). *The Quarterly Review of Biology, 68*(4), 533–546.
Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*(4), 579–600.
Perrin, J. (1913). *Les Atomes* (D. Ll. Hammick, Trans.). Woodbridge: Ox Bow Press.
Reichenbach, H. (1956). *The direction of time*. Berkeley: University of California Press.
Schupbach, J. N. (Forthcoming). Robustness analysis as explanatory reasoning. *British Journal for the Philosophy of Science*.
Stolarz-Fantino, S., Fantino, E., Zizzo, D. J., & Wen, J. (2003). The conjunction effect: New evidence for robustness. *The American Journal of Psychology, 116*(1), 15–34.
Weisberg, M., & Reisman, K. (2008). The robust Volterra principle. *Philosophy of Science, 75*(1), 106–131.
Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 125–163). New York: Jossey-Bass. Page references are to the version reprinted in Wimsatt (2007).
Wimsatt, W. C. (1994). The ontology of complex systems: Levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy, 24*(Suppl. 1), 207–274. Page references are to the version reprinted in Wimsatt (2007).
Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Cambridge: Harvard University Press.
Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13*(2), 219–240.

Part VII
Fiction, Representation and Explanation

Why Does Water Boil? Fictions in Scientific Explanation

Sorin Bangu

1 Introduction

We are all familiar with boiling water. But, is this process fully accounted for scientifically? The question is surprising; while there are still many unexplained physical phenomena, this one never seemed to be among them. Yet an authority in the field of thermal physics, David Ruelle (1991, 123–4), writes the following:

> So, here is a problem for the theoretical physicist: prove that as you raise or lower the temperature of water you have phase transitions to water vapor or to ice. Now, that's a tall order! We are far from having such a proof. In fact there is not a single type of atom or molecule for which we can mathematically prove that it will crystallize at low temperature. These problems are just too hard for us.

Although this is said in a physics popularization book, the main statement of the quote is literally true. As we'll see in a moment, this type of problem – understanding how phase transitions occur[1] – can be spelled out in more precise terms, and it will turn out to be 'hard' indeed. The worries are of a conceptual nature, and remain to be contemplated even today, after the serious technical-mathematical difficulties underlying them have been overcome. This puzzle drew the attention of philosophers and philosophically-minded physicists, and several papers have already discussed it in considerable detail; it is now common to refer to

[1] Phase transitions are in itself a fascinating topic, as they appear everywhere in science, from cosmology to economics. There are several types of phase transitions, but here I will be concerned only with first-order ones.

S. Bangu (✉)
Department of Philosophy, University of Bergen, 12/13 Sydnesplass, Bergen 5007, Norway
e-mail: sorin.bangu@fof.uib.no

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_24

it as the 'paradox' of phase transitions.[2] One direction this literature followed was to examine the bearing of this paradox on the philosophical debate on emergence and reduction. Here I will have virtually nothing to say about these notions[3], as my primary concern is a different, yet related issue: the relevance of the paradox for understanding the role of fictions in scientific explanation.

There has been a lot of work devoted to this issue lately, since many philosophers of science agree with M. Suarez's assessment, that this is "one of the main and most controversial roles that fictional assumptions may play" (Suarez 2009, 7). Needless to say, fictionalism in philosophy is an old and well-covered theme; in philosophy of science, it goes back to Hans Vaihinger and his 1911 (1924, 1952) book *The Philosophy of 'As If'*. Vaihinger argued, among other things, for a negative answer to the central question of interest here – as to whether a fiction can explain. He thinks of fictions rather broadly, as something which "falsifies reality" (1952, 88), and believes that "... the fiction induces only an illusion of understanding" (1952, xv), and that "[F]iction (...) does not create real knowledge" (1952, 88). This skepticism is in agreement with our intuitions – it's not a good explanation of the presents under the Christmas tree that Santa Claus brought them – and the negative answer has been the received view for several decades.[4] The discussion of the role of fictions in science has been revived in the early 1990s by Arthur Fine's seminal paper *Fictionalism* (Fine 1993), and the literature has since grown exponentially. Today the landscape is rather varied[5], and one can find authors who are more open toward a positive answer.[6] In what follows, I will join this group, although my reasons for holding this view are different from theirs.[7] In a nutshell, the position I advocate here is that in certain cases in science the acceptance of fictions among the explanans is justified indeed, since they are required by the special nature of the explananda: when the explananda are a certain kind of ineliminable fictions themselves, then there should be no surprise that the scientists find the use of fictional explanans entirely acceptable.[8]

The paper proceeds as follows. In Sect. 2, I will present the (so-called) paradox of phase transitions and two approaches to it. Then, in Sect. 3, I will argue that what I take to be the most satisfactory solution to this philosophical conundrum can

[2]For a most recent example, see Shech 2013. Other contributions include Humphreys 1997; Liu 1999; 2001; Callender 2001; Callender and Menon 2013; Batterman 2005; Bangu 2009; 2011; Butterfield 2011; Norton 2012; Morrison 2012; Kadanoff 2013.

[3]My take on this is in Bangu (forthcoming); see also Bangu 2009 and 2011, from which I borrow the general presentation of the conceptual issues involving phase transitions in Sect. 2.

[4]As Bokulich 2009 notes, Hempel's covering-law theory of explanation and Salmon's causal-ontic account are hostile to the idea that fictions can appear among explanans.

[5]See Suarez's 2009 excellent collection.

[6]See, for instance, the essays by Morrison, Elgin, Bokulich and Winsberg in Suarez 2009.

[7]For reasons of space, however, I can't compare my approach to other approaches.

[8]Whether these are the only cases is a larger question which I can't take up here, although I suspect the answer is negative.

be naturally integrated into a larger discussion about the role played by fictions in scientific explanation. The fiction under scrutiny here is, as we'll see, a physically impossible object: a statistical mechanical system with an infinite number of degrees of freedom (i.e., consisting of an infinite number of particles (molecules)). I will compare this use of the infinite fictional system with another case – from electrostatics – in which we also derive results by assuming infinite values for certain physical quantities.

2 The 'Paradox' of Phase Transitions

The fact that substances change their aggregation state, often abruptly, was of course known since times immemorial; its systematic study, however, began only in the nineteenth century.[9] Classical thermodynamics offered a macro-level perspective on the process of phase change, describing it as the crossing of a 'coexistence line' (eg. OC, in Fig. 1 below) in a schematic 'phase diagram'.

As is well known, classical thermodynamics has developed a rather sophisticated conceptual apparatus to describe these phenomena.[10] One typically defines the (Gibbs) 'free energy' of the system as $G = H - TS$, the difference between the enthalpy of the system and the product of (absolute) temperature and thermodynamical entropy. See Fig. 2; the dotted vertical line separates the two phases.

The crossing of a coexistence line (corresponding to, say, vaporization) takes place when function G displays special mathematical properties – that is, when G, more generally characterized as a 'thermodynamic potential', features a kink, or a

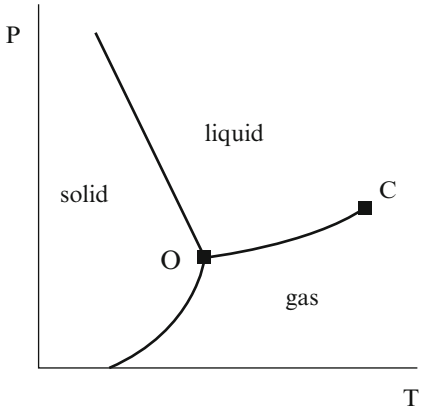


Fig. 1 The three phases are separated by lines connected at *point O*, where all phases coexist (this is the 'triple' point.) Beyond *point C* (the 'critical' point) the liquid and the gas phases are indistinguishable; the substance is a 'fluid'

[9] See Andrews 1869. Major contributors are, among others, by R. Clausius, J. D. Van der Waals and J. C. Maxwell.

[10] I follow Zemansky's 1968 classical textbook treatment.

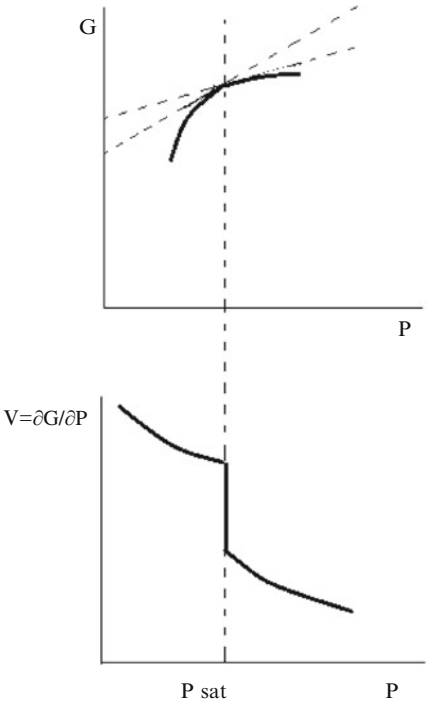


Fig. 2 G is Gibbs free energy, V is volume and P is pressure. The phase change is represented as the kink on G's graph (*top*), or as a discontinuity in volume (*bottom*). The pressure here is not the critical pressure, but the value of the saturated vapor pressure. The temperature is less than the critical temperature. After Stanley 1971, 31

sharp corner. Mathematically, the transition is described as a 'singularity', or 'non-analyticity'. Equivalently, a (gas–liquid) first-order transition takes place when the volume (a first derivative of G) changes discontinuously.

However, the branch of physics in charge of *explaining* these phenomena today is not thermodynamics, but its conceptual successor, the more 'fundamental' theory of statistical mechanics. This theory is supposed to recover (and sometimes correct) the phenomenological results systematized by classical thermodynamics.[11] Statistical mechanics has been generally successful in doing this, and one would then expect that it would be able to deal with phase changes too – by showing that the statistical mechanical counterparts of the thermodynamical potentials (such as G) feature non-analyticities. Now, as Ruelle pointed out above, *this* is a tall order.

The root of the difficulty is the way in which the two theoretical frameworks, the macroscopic (classical thermodynamics) and the microscopic (statistical mechanics) get connected. Thermodynamical quantities, free energies included, are typically expressed in statistical mechanics in terms of the partition function (typically denoted by 'Z') (see Reif 1965, 214–16 for details.) But Z is a finite sum of analytic ('smooth') functions, and thus a smooth function itself, so it can't, as a matter of pure mathematics, harbor any non-analyticities. It follows that the

[11] Hence the widespread idea that this is a case of theory reduction. See Nagel 1961.

thermodynamic potentials like G can't feature singularities, and the discontinuities (in the volume[12]) can't be demonstrated either. No wonder then that many became tempted to present this situation as a failure of reduction; and, if emergence is (loosely) defined as failure of reduction, one can then claim that vaporization is an example of an emergent phenomenon. The derivation of a singularity is impossible in principle, and not just due to computational complexity, hence this case would qualify as more than (merely) epistemological ('weak') emergence, but rather as ontological ('strong') emergence. Several notable physicists and philosophers consider these cases "paradigms of emergent behavior" (Lebowitz 1999, S346), and maintain that "the existence of phase transitions shows that we have to be careful when we adopt a reductionist approach. Phase transitions correspond to emerging properties." (Prigogine 1997, 45) C. Liu calls phase transitions "truly emergent properties" (1999, S92).[13]

Interestingly, all these pronouncements came in full awareness of the fact that the physicists were able, eventually, to address the singularity problem. The solution was a remarkable achievement[14], made possible, in essence, by considering a *fictional* system, composed of an infinite number of particles – that is, by taking the 'thermodynamic limit'[15]; then one can derive singularities in the statistical mechanical versions of the relevant thermodynamic potentials.

The 'paradox' can now be stated more clearly. Vaporization seems an unremarkable, mundane phenomenon; it takes place every morning in the finite amount of water in your tea kettle. And yet, as Leo Kadanoff said, "the existence of a phase transition requires an infinite system. No phase transitions occur in systems with a finite number of degrees of freedom." (2000, 238) In other words, when it comes to a finite, real system, statistical mechanics can't tell us why it will eventually turn into vapors; the derivation works only if a fictional infinite system is considered, which thus seems to play an ineliminable explanatory role. But since infinite systems don't exist, one may be tempted to conclude that we haven't really answered the immemorial question 'why does water boil?'

One important query cropping up at this point is what it would take to come up with an unobjectionable solution to this problem – which I will call the 'singularity', or 'discontinuity' problem. On the face of things, the most direct way to deal with it would be to present a derivation of the non-analyticities *within* statistical mechanics and *without* involving the fictional infinite system. Another option would be to keep

[12]The same holds for others 'signatures' of a phase change, for instance the divergence of heat capacity.

[13]See also Humphreys 1997.

[14]Onsager, and Yang and Lee are among the key-contributions on a long list that includes even Einstein.

[15]Further significant constraints are also imposed on this system, one of them being that the ratio between the number of particles in it and the volume it occupies is finite. See Liu and Emch 2002 for more technical details, and Sklar's (1993, 78–81) account of the reasons for which taking the thermodynamic limit is so useful: among other things, it establishes the equivalence of ensembles, deals with system's boundaries, and gets rid of the effect of fluctuations.

the fictional element in play, thus recognizing its ineliminability, but neutralize the worries prompted by it. Both ideas have been tried, and I will discuss them in turn. Neither works perfectly, but the second is more promising and, I will argue, connects naturally with the issue of explanatory fictionalism.

As is clear, the first strategy would involve a novel statistical mechanical approach to singularities. Some physicists have made remarkable efforts in this direction; see the work of Franzosi, Pettini and Spinelli 2000, for a sample. The key-idea they pursued was to relate the singularities (of the micro-canonical entropy) to the thermodynamical phase transitions. The overall aim of this approach is to demonstrate that there are non-analyticities in the entropy corresponding to a change in the topology of configuration space.[16] Some philosophers are quite enthusiastic about this idea; Callender and Menon 2013, for example, believe that "it is clear that the microcanonical ensemble does exhibit singularities even in the finite particle case and that there is a *plausible* research program attempting to understand phase transitions in terms of these singularities." (2013, 217; my italics)

As I pointed out in my (forthcoming), I suspect that this optimism is somewhat exaggerated. For one thing, Callender and Menon themselves are aware of several "open questions" (2013, 217) still to be answered by this approach. They wonder, for example, "what topological criteria will be necessary and sufficient to define phase transitions, if any such criteria can be found." (2013, 217) The worry is surely pertinent, since in a later paper (not cited in Callender and Menon 2013), the same Franzosi and Pettini 2004 explicitly warn that the theorem they proved shows that "...a topology change (...) is a *necessary* condition for a phase transition to take place at the corresponding energy or temperature value" (italics in original), while "...the converse of our Theorem is not true. There is not a one-to-one correspondence between phase transitions and topology changes; in fact, there are smooth, confining, and finite-range potentials (...) with even a very large number of critical points, and thus many changes in the topology (...) but with no phase transition. Therefore, an open problem is that of *sufficiency* conditions, that is, to determine which kinds of topology changes can entail the appearance of a [phase transition]." (italics in original). It thus seems that Callender and Menon's confidence that "statistical mechanics might well have the resources to adequately represent these discontinuities without having to advert to the thermodynamic limit" (2013, 217) is still far from being vindicated.[17]

[16]Here is how Franzosi et al. 2000, 2774 describe their central idea: "a major topology change (...) is at the origin of the phase transition in the model considered." Furthermore: "suitable topology changes of equipotential sub-manifolds of configuration space can entail thermodynamic phase transitions.(...) The method we use, though applied here to a particular model, is of general validity and it is of prospective interest to the study of phase transitions in those systems that challenge the conventional approaches, as might be the case of finite systems."

[17]Callender and Menon 2013 also discuss other approaches in addition to the topological idea. One is the 'back-bending' in the microcanonical caloric curve (section 3.1.1), the other is the perpendicular distribution of zeros (section 3.1.2). These approaches, however, fare no better than the topological one just discussed; they point out that "[p]robably none of the definitions

3 Honest Fictions: Measurements and Their Representation

Since the first way out is not very promising, we need to examine the other solution. It consists in objecting to importing the thermodynamical definition of a phase transition (as a singularity) within statistical mechanics. This is so because, as Callender memorably put it, we should not take thermodynamics 'too seriously': "After all, the fact that thermodynamics treats phase transitions as singularities does not imply that statistical mechanics must too." (2001, 550) On the face of things, this does dissolve the paradox: if phase changes are represented in statistical mechanics in a different way – that is, if the mathematical property we have to explain / derive (the singularity) is replaced by something else – then one can hope, even expect, that the troubles created by the fictional infinite system will go away too.

As is evident, this strategy immediately raises a new question: if not as singularities, then . . . how else to represent phase transitions in statistical mechanics? As a philosopher, Callender is of course under no obligation to answer this question (and, to be sure, he doesn't). From the perspective of a physicist, however, the query is more pressing. To follow up this line of thinking would amount to making substantive changes in the current formulation of thermal physics, which would involve writing down a different theory of phase transitions in statistical mechanics. But the physicists, it seems, are in no rush to do this, and thus their acceptance of the situation is in need of further examination.

Although this solution leaves something to be desired, it puts us on the right track: we should pay more attention to what our explanandum is. This line of inquiry is just natural, in so far as, on reflection, this case is definitely not unusual in science. In fact, infinite quantities are often needed in order to work out the derivations in the textbooks; virtually all scientific disciplines contain references to infinitely deep oceans, infinitely long wires, infinitely large populations, etc. I will now briefly present one such simple textbook example – from electrostatics – and then return to the phase transitions case. I hope the comparison will be illuminating, and will reveal an interesting facet of the issue of how fictional infinite objects come to play a role in an explanation.

Consider a uniformly charged disk of radius R, and various points on its axis, located at different distances above its center, on the same side. The disk is uniformly charged and the charge density (per unit area) is constant, σ. Now, suppose that one measures the electrical field at some of these points and finds that it remains constant, with a value given by

$$E = \sigma/2\varepsilon_0,$$

provide necessary and sufficient conditions for a phase transition that overlaps perfectly with thermodynamic phase transitions". They also add, but without further clarifications of the claim, that "[t]hat, however, is okay, for thermodynamics itself does not neatly characterize all the ways in which macrostates can change in an 'abrupt' way." (2013, 210)

where ε_0 is constant ($\varepsilon_0 =$ the permittivity of space). The constancy of E is prima facie puzzling, and we would like an explanation of it. The explanation usually proceeds by deriving the expression of the electric field E_H at an arbitrary point H situated at height h on the disk's axis[18]:

$$E_H = (\sigma/2\varepsilon_0)\left[1 - \left(1 + (R/h)^2\right)^{-1/2}\right]$$

As is clear, to derive the constancy of the electrical field at various elevations h (as $\sigma/2\varepsilon_0$) we have to let $R \to \infty$. But, since the disk is not infinite, something is going amiss here; should we perhaps begin worrying about the 'paradox' of the constant electric field now? Facing this puzzle, our first reaction, I suspect, will be to go back and check the measurements of the electric field. Do they *truly* show that the field is constant? We thus look for errors – due, for instance, to the calibration of the instrument, the presence of 'noise', etc. Importantly, even if no such issues are found, the instruments may still indicate a constant field, and this is so because their accuracy is *in principle* limited: there can surely exist two heights, h and $h + \xi$, for which no discrepancy in the measured values of the field is detected, in so far as it falls below the experimental error of the apparatus.

The point of discussing this aspect is to suggest that we should ask the same question in the case of phase changes: is their phenomenology such that we must represent them as mathematical points, or singularities? This is the key-question here, which seemingly got forgotten in the debate. It needs to be brought back to the forefront – while keeping in mind that Liu and Callender have already answered it in the negative: "we don't actually *measure* perfect singularities" (Callender 2001, 550; my emphasis. See also Liu 1999; 2001). This means that the kinks in the thermodynamic potentials of finite, real systems are not observable; in the case of a real, finite system "the transition is neither 'smooth' nor 'singular'" (Liu 1999, S103). Then, in a sense, we need not worry about the singularities, since they are in fact *added* to the isotherms. But singularities are, recall, our explananda, and yet, they are, as Liu called them, "artifacts" (1999, S104) of thermodynamics – or "fictions" (2001, S336). A comparison of Figs. 2 and 3 below is enlightening.

Unlike the schematic Fig. 2, Fig. 3 depicts an actual, and typical, plotting of a laboratory measurement. We can see a change of phase for the substance (yttrium) occurring as the pressure increases within the range 95–100 GPa. The phase called 'hR24' is represented by square data-points, while the other phase, 'C2/m', by triangle data-points. A discontinuity, or a marked 'drop' in the volume[19], appears as the pressure reaches 97 GPa. It is crucial to note that this happens within a *region* (encircled on the Fig. 3); its representation in Fig. 2 is not a direct reflection of what is actually measured, but is required by the thermodynamic definition.

[18]For a derivation, see Halliday, Resnick, and Walker 2011, sect. 22–7.

[19]In the study, the volume change is measured as −2.6 % of the volume.

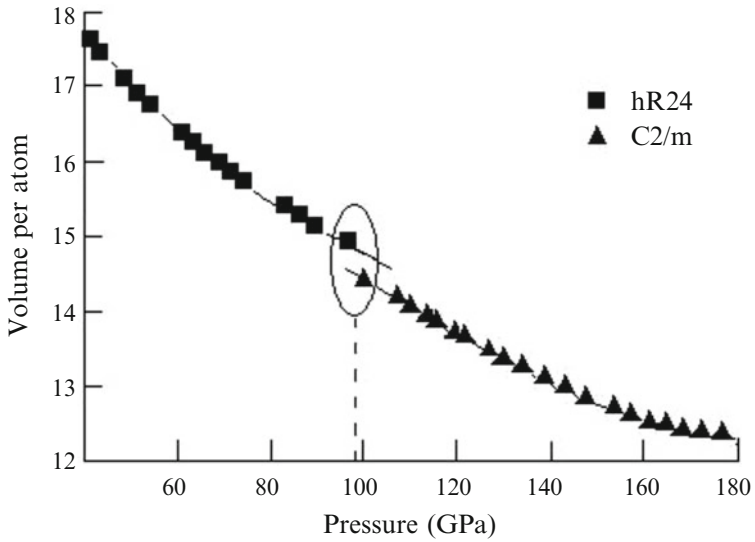


Fig. 3 A phase-change of yttrium (a rare earth element) at room temperature (Drawn after part (a) of Fig. 3 in Samudrala et al. 2012, 3). The encircling, and the dotted vertical line indicating the approximate pressure of 97 GPa, have been added to the original figure

The analogy between the electrostatic and the thermodynamical cases is now hopefully transparent. In the electrostatic case, we need to appeal to an infinitely large system (the disk) because we wanted to explain the constancy of the field. But, after checking the measurements, it turned out that measuring a perfectly constant field was illusory – or, if we did in fact record this, this was due to inherent limitations on the accuracy of the instruments. The same holds in the case of the thermodynamical singularities: the 'kink' in Fig. 2-top, corresponding to the vertical drop, is an addition to the graph; it is not measured, but an artifact of the representation of the measurement. It is thus crucial to note that (i) in both the electrostatic and the thermal cases, the explanandum itself is fictional and, more importantly, that (ii) it is the attempt to account for a fictional explanandum (the constant electrical field, the singularity) that demands a fictional explanans (the infinitely large disk, the infinite statistical-mechanical system). Now we can see how Callender's and Liu's (second) approach to the paradox discussed above bears on the issue of explanatory fictionalism: in explanatory terms, we see that the appeal to 'fictional' explanans (the infinite system) is needed because of the 'fictional' nature of explananda (the singularities).[20] Once we see things in this light, no special worries about the explanatory role of fictions should arise.

[20] As Liu put it: "[T]he idealization in SM [statistical mechanics] ... is required by the idealization in TD [thermodynamics]" (1999, S103)

This should be elaborated further, however, since calling the singularities (and the infinite systems) 'fictions' can be easily misinterpreted. As is hopefully evident by now, they are not fictions in the usual sense of the word – that is, in the same sense in which Santa Claus is a fiction. Unlike Santa, singularities fall in a special category of fictions 'concerned' with the truth, to adapt Winsberg's 2009 apt phrase; these are 'honest' fictions. To call the explananda 'fictional' is, in this context, *not* to downgrade them (epistemically and ontologically), but to recognize that they have this status since the scientists' epistemic condition is subject to ineliminable constraints: the accuracy of measurements is *in principle* limited.[21] A related, and perhaps helpful way in which one may understand the talk of 'honesty' here is by analogy with the financial transactions in the famous MONOPOLY board game: there is no problem to pay with fictional money precisely because the real-estate itself is fictional – and the analogy is: there should be no problem to invoke fictional explanans precisely because the explananda are fictional too. Naturally, worries appear when the real-estate is *not* fictional, and yet one pays for it with fictional money (as recent events, especially in the US, have shown).

Although we examined only two examples, their lesson is more general: we have identified a certain type of scientific explanation in which the appeal to fictional explanans is acceptable, and these are cases in which the explananda themselves are fictions (to stress: 'honest' fictions). In fact, both the explananda and the explanans are ineliminable fictions, introduced on the basis of broad and well-motivated theoretical and experimental considerations.

This completes the sketch of what I see as a promising angle on the fundamental question about the explanatory status of fictions. The novelty of the approach sketched here consists in a change of perspective. The focus of this debate has always been the explanans, and the central worry was that admitting fictions among the explanans destroys the credibility of a theoretical account. Here, drawing on the (second) solution of the paradox discussed above, I proposed to turn this picture upside-down and concentrate on the explananda, by asking: aren't they (honest) fictions as well? If so, there's no need to worry that the explanans are (honest) fictions too.

4 Conclusion

At one extreme, anything can be 'explained' if we don't impose restrictions on the nature of the explanans (the Santa story is, again, a case in point, as are, more generally, all religious 'explanations'). Such constraints must therefore be imposed,

[21] There are no data-points for 96.36 GPa (to take an arbitrary value) and even if there were, an increasing resolution would still reveal *regions* of transition. For more on the role of measurements, and how scientists shape the data-points into 'phenomena' (in the well-known sense of Bogen and Woodward), see Bangu 2009.

but we have to be careful with their strictness, since too severe ones would render doing science virtually impossible. Scientists regularly appeal in explanations to departures from the literal truth, i.e., to fictions; ever since Galileo have we learned to live with the idea that certain fictions (frictionless surfaces, point-particles, etc.) will always be ingredients of the explanans. As many authors have noted, the acceptance of these types of fictions is grounded in their expediency and their eliminability (at least in principle). While I agree with this view, my suggestion here was to investigate concrete cases in which the appearance of fictions among the explanans is justified not only by invoking ineliminability considerations, but also by pointing out that they are required by the special nature of the explananda: when the explananda are (honest) fictions, the scientists find the use of (honest) fictional explanans entirely acceptable too.

Acknowledgement I thank the three anonymous referees for their criticisms of a previous version of the paper.

Bibliography

Andrews, T. (1869). On the gaseous and liquid states of matter. *Philosophical Transactions of the Royal Society of London, 159*, 575–590.

Bangu, S. (2009). Understanding thermodynamic singularities. Phase transitions, data and phenomena. *Philosophy of Science, 76*(4), 488–505.

Bangu, S. (2011). Bridge laws in inter-theoretic relations. *Philosophy of Science, 78*(5), 1108–1119.

Bangu, S. (forthcoming). Neither weak, nor strong? Emergence and functional reduction. In M. Morrison & B. Falkenburg (Eds.) *Why more is different. philosophical issues in condensed matter physics and complex systems*. Berlin: Springer.

Batterman, R. (2005). Critical phenomena and breaking drops: Infinite idealizations in physics. *Studies in History and Philosophy of Modern Physics, 36*, 225–244.

Bokulich, A. (2009). Explanatory fictions (pp. 91–109) in Suarez (2009).

Butterfield, J. (2011). Less is different: Emergence and reduction reconciled. *Foundations of Physics, 41*, 1065–1135.

Callender, C. (2001). Taking thermodynamics too seriously. *Studies in History and Philosophy of Modern Physics, 32*, 539–553.

Callender, C., & Menon, T. (2013). Turn and face the strange. Ch-ch-changes: Philosophical questions raised by phase transitions. In R. W. Batterman (Ed.), *Oxford handbook for the philosophy of physics*. New York: Oxford University Press.

Fine, A. (1993). Fictionalism. *Midwest Studies in Philosophy, XVIII*, 1–18.

Franzosi, R., & Pettini, M. (2004). Theorem on the origin of phase transitions. *Physical Review Letters, 92*, 060601.

Franzosi, R., Pettini, M., & Spinelli, L. (2000). Topology and phase transitions: Paradigmatic evidence. *Physical Review Letters, 84*, 2774–2777.

Halliday, D., Resnick, R., & Walker, J. (2011). *Fundamentals of physics* (9th ed.). New York: Wiley.

Humphreys, P. (1997). How properties emerge. *Philosophy of Science, 64*, 1–17.

Kadanoff, L. (2000). *Statistical physics*. Singapore: World Scientific.

Kadanoff, L. (2013). Theories of matter: Infinities and renormalization. In R. W. Batterman (Ed.), *Oxford handbook for the philosophy of physics*. New York: Oxford University Press.

Lebowitz, J. L. (1999). Statistical mechanics: A selective review of two central issues. *Reviews of Modern Physics, 71*, S346–S347.
Liu, C. (1999). Explaining the emergence of cooperative phenomena. *Philosophy of Science, 66*, S92–S106.
Liu, C. (2001). Infinite systems in SM explanations: Thermodynamic limit, renormalization (semi-) groups, and irreversibility. *Philosophy of Science, 68*, S325–S344.
Liu, C., & Emch, G. (2002). *The logic of thermo-statistical physics*. Heidelberg: Springer.
Morrison, M. (2012). Emergent physics and micro-ontology. *Philosophy of Science, 79*, 141–166.
Nagel, E. (1961). *The structure of science*. New York: Harcourt, Brace and World.
Norton, J. (2012). Approximation and idealization: Why the difference matters. *Philosophy of Science, 79*(2), 207–232.
Prigogine, I. (1997). *End of certainty*. New York: The Free Press.
Reif, F. (1965). *Statistical and thermal physics*. New York: McGraw-Hill.
Ruelle, D. (1991). *Chance and chaos*. Princeton: Princeton University Press.
Samudrala, G., Tsoi, G. M., & Vohra, Y. K. (2012). Structural phase transitions in yttrium under ultrahigh pressures. *Journal of Physics. Condensed Matter, 24*, 362201.
Shech, E. (2013). What is the paradox of phase transitions? *Philosophy of Science, 80*(5), 1170–1181.
Sklar, L. (1993). *Physics and chance*. Cambridge: Cambridge University Press.
Stanley, E. H. (1971). *Introduction to phase transitions and critical phenomena*. Oxford: Oxford University Press.
Suarez, M. (Ed.). (2009). *Fictions in science*. London: Routledge.
Vaihinger, H. (1952). *The philosophy of 'as if'*. (trans: Ogden, C. K.). London: Lund Humphries. (First published, 1911; Translated by C. K. Ogden and printed in 1924 by Kegan Paul).
Winsberg, E. (2009). A function for fictions: Expanding the scope of science (pp. 179–192) in Suarez (2009).
Zemansky, M. W. (1968). *Heat and thermodynamics: An intermediate textbook* (5th ed.). New York/Tokyo: McGraw-Hill.

Scientific Representation, Denotation, and Fictional Entities

Mauricio Suárez

1 Introduction

The influential Denotation-Demonstration-Interpretation (DDI) account of representation was developed in a short pioneering paper by RIG Hughes (1997). My purpose in this paper is to assess the DDI model in light of present day interest on the nature of fictional entities in science. I argue that the DDI model faces an insurmountable difficulty in dealing with such entities. However, an extended version of the DDI account may accommodate fictional entities. While the resulting account is more complex, this may just reflect the complexity of representation itself. In addition the extended version is clearer with respect to the key question regarding the deflationary nature of representation, since it makes it patent that representation is not a relation between its source and target systems, but a functional property of models within a representational practice.

In the first section, I review the original DDI proposal, emphasizing the role that the relation of denotation plays in this proposal. In the second section, I discuss and emphasize the deflationary nature of the DDI account. In Sect. 3 I briefly review some examples of scientific fictions, particularly Maxwell's vortex model of the ether, and show that the DDI account fails to accommodate them. I argue instead for a weakening of the denotation and interpretation relations into correlative functional notions. The conclusion emphasizes the deflationary nature of the suitably extended

M. Suárez (✉)
Institute of Philosophy, School of Advanced Study, London University, Senate House, Malet Street, London WC1E 7HU, UK

Department of Logic and Philosophy of Science, Faculty of Philosophy, Complutense University of Madrid, Madrid 28040, Spain
e-mail: msuarez@filos.ucm.es

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_25

account, and how it reveals that representation is not a relation *per se*, although it can be instantiated by means of certain relations in certain contexts.

2 The Denotation-Demonstration-Interpretation (DDI) Account

The DDI account of representation was introduced by RIG Hughes in his now classic paper (Hughes 1997). In order to outline and assess the DDI account we need first to fix some neutral terminology. We shall say that, in model-building science, a model source A typically represents a target B. This terminology implies no constraints on what types of objects A and B may be: These may be concrete or abstract, physical or mathematical, real or imaginary. Neither does it preclude the standard view according to which any scientific model must have a target in the real world and represent it via relations that hold between the properties of both source and target. Indeed, as discussed below, the standard view is constitutive of representation on most substantive accounts, which take representation to be a relation – and hence take both relata to be real. Yet, the terminology also leaves room for other views that do not require sources or targets (or both) to be real, and hence do not require representation to be a relation.

In other words many types of objects can play the role of representational sources – from concrete physical objects and diagrams to abstract mathematical structures or laws. And in addition, an indefinite number of different sources may represent one and only one target. Thus a concrete array of small balls carefully strung together by means of wires and Kepler's mathematical laws can both meaningfully represent the solar system, albeit to very different degrees of accuracy. Similarly for targets, the variation here can be large. Some models represent concrete physical systems and their dynamical evolution, such as the solar system; other models represent more general phenomena, or effects, such as the Ising model for phase transitions; or abstract properties, such as the second law of thermodynamics, which represents entropy as necessarily increasing in closed systems.

What do all these instances of 'scientific representation' have in common? This has not been an easy question to answer and there are a number of different proposals. We may, however, classify the proposals available in roughly two different kinds: substantive and deflationary.[1] Substantive approaches answer the question in terms of the properties of sources and targets – and their relation – that constitutes representation. By contrast, on a 'deflationary' approach there is in fact no substantive or explanatory property or relation that constitutes representation. What is rather in common between the different cases of representation is the

[1]For an elaboration of the distinction between deflationary and substantive accounts, see Suárez (2010) and, particularly Suárez (2015) of which the above is an abbreviated version.

cognitive role or function that sources play vis. a vis. their targets – i.e. the uses that they are put to by agents towards their specific goals in their particular contexts of inquiry. And that's that. There are no further conditions lurking, as it were, in the background. Hughes' DDI account is, at least prima facie, an approach to representation of this deflationary kind.

The DDI account takes it that scientific modelling is a hybrid notion, containing both elements characteristic of a relation and others more of a piece with an activity. On this view, a source A represents a target B when the following three conditions are met: (i) The source denotes the target; (ii) A demonstration is carried out on the model; and (iii) The results of this demonstration are then interpreted in terms of the target.

Hughes vividly illustrates these elements by reference to the model that Galileo introduces in the Third Day of his *Discourses Concerning Two New Sciences*. Galileo there describes a kinematical problem in geometrical terms, solves the problem in geometry, and then applies the solution back to the original kinematical problem. In particular he concludes that the space *s* traversed by a body in uniform motion with constant velocity in a given interval *t* is equal to that traversed by a uniformly accelerated body initially at rest, provided that the final speed of the accelerating body is twice that of the body in uniform constant motion. In terms of the DDI model, he reasons as follows. First, the kinematical situation must be described by means of a geometrical diagram that therefore denotes it (Fig. 1). Thus Galileo denotes the time *t* that the body takes to traverse the space *s* by means of the segment AB of a line, and the speed of the body at any instant of the interval t by

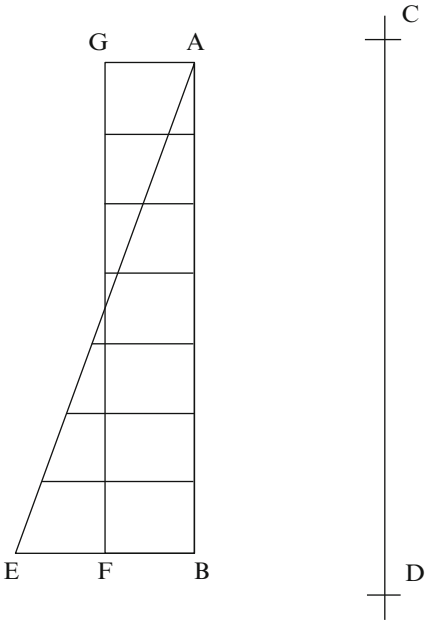


Fig. 1 Galileo's geometrical model

another segment of a line perpendicular to the first line. Thus AC denotes the speed of the body at A and BD the speed of the body at B. Second, a demonstration must be carried out on the diagram. Galileo demonstrates that the area of a rectangular shape ABCD is identical to the area of a triangle ABD' where D' is twice the value of D. Finally, the result is interpreted back in the terms of the original kinematical problem, by conceiving the overall area covered as the space traversed by the body in its motion over the *t* interval. Thus Galileo concludes that the time *t* that a body in uniform motion takes to traverse *s* is identical to the time taken by a body uniformly accelerated. QED. The three stages in Galileo's reasoning coincide neatly with the denotation, demonstration and interpretation stages in the DDI account.

3 The Deflationary Nature of the DDI Model

Hughes presents the DDI account in a rough and ready way as a deflationary approach to representation because he explicitly refrains from postulating necessary conditions in terms of robust relations between sources and targets (1997, p. 329): "Let me forestall possible misunderstandings. I am not arguing that denotation, demonstration, and interpretation constitute a set of speech acts individually necessary and jointly sufficient for an act of theoretical representation to take place". This is to deny that the DDI account provides us with a substantive explanatory property since it does not even provide necessary and sufficient conditions.[2] Nevertheless, there are a few features of the DDI account that may lead us to question the strength of its commitment to deflationism. These features all follow from the surprising appeal to denotation, which is commonly understood as a substantive relation between the denoting sign and the denoted object.[3]

Hughes' account may be summarized in a schema (Fig. 2) which should not hide its hybrid nature. Denotation is a relation between a source and a target; while demonstration and interpretation seem best understood as activities on the part of an interpreter/user. There is, of course, an activity of denoting – but this is commonly understood to either establish a relation, or ride upon an already established one. In other words, we may not use A to denote B without ipso facto establishing a relation of denotation between A and B. The relation substantially informs the notion of representation at play, as revealed by our use of the language. For instance, we speak of the geometrical diagram as in itself denoting the kinematical problem, independently of any activity carried out by Galileo, as if the relation of denotation was entirely independent of anything we can actually do or not with it. There is at least prima facie a question here regarding the nature of the relation that informs this conception of representation. The contrast is great with the notion of demonstration, which can only be conceived as a piece of reasoning carried out by someone entirely within the 'space of reasons' provided by the model source. It seems hopeless to

[2] For a more detailed discussion of this point, see Suárez (2015).

[3] Goodman (1976); see also Elgin (1996, 2009).

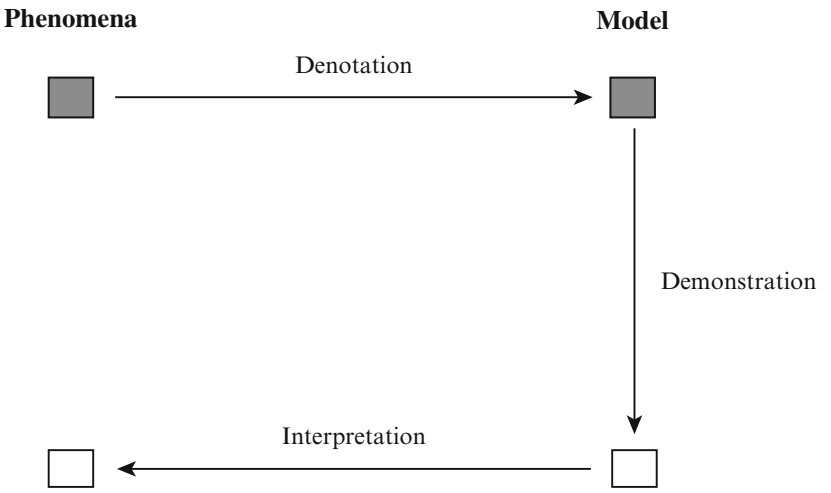


Fig. 2 The DDI model

attempt to interpret this as a relation, since at this stage of the modelling process, the target may not taken into consideration at all. So, on the DDI account, in order for the geometrical model to represent (for us) the kinematical situation, we must carry out Galileo's demonstration ourselves. It would not seem to be true that "there is a demonstration out there, waiting for us to apprehend it". Here, by contrast with the denotation part, the activity itself is constitutive of representation, and there is no relation that may stand in its place.

The third element in the DDI account appears less clear-cut. In model theory, of instance, the notion of 'interpretation' may be understood as a relation[4]: It is a function mapping the elements of the language into a domain of independent entities endowed with their own properties. Hence, take a set of sentences in some particular language; the 'interpretative mapping', on this account, provides them with a 'semantics' under which they may be said to be true or false. But it is doubtful that this is the same 'interpretation' that is involved in the DDI account, since to the extent that the model source contains sentences at all, they already come fully interpreted in terms of the model itself. It seems more appropriate to think of it as an instance of 'application': it applies the model source to the target in order to derive results of interest regarding the target itself. Now, there is no doubt that the application of the model is constrained by the relation of denotation established in the first stage of the DDI account, but it also brings a large degree of freedom in two respects at least. Firstly, the denotation relation by itself does not stipulate which parts of the target object correspond to which parts of the source object, and there is always plenty of leeway at this point. In the Galileo example

[4]For instance see Chang and Kleiser (1990, p. 20ff).

the mere fact that the geometrical diagram denotes the kinematical situation does not settle which parts of the diagram stand for which parts of the kinematics. But more importantly the mere fact of denotation does not determine how the source is to be conceived in the first place, i.e. how it is to be divided into parts that can then be related to the target. And it is, however, clear that the application of the source to the target does require a partition of the source into relevant parts and properties (a "structure"), and the relating of such "structure" to a similar "structure" of parts and properties in the target. Thus in Galileo's modelling example, the geometrical diagram must clearly distinguish vertical and horizontal lines at every point, and the area therein comprised. Similarly the kinematical problem must clearly identify time intervals, speed of motion at every instant, and constant or accelerated motion across the interval. Etc.

In other words, 'interpretation' requires at least two types of activity on the part of the modellers. First, it requires the ascribing of some structure to the source and target objects, by judiciously partitioning them into an appropriate set of features and their properties. Second, it calls for a mapping of the elements of the source structure onto some corresponding parts and properties of the target, again under some suitable partition, which is typically ascribed on pragmatic grounds.[5] Both steps ('partitioning' and 'mapping') are activities within the modelling practice without which interpretation is impossible. However, only mapping issues in a sort of relation akin to denotation between (elements of) the source and (elements of) the target. Therefore, on the DDI account, modelling is a hybrid of a relation (denotation, mapping), and a number of activities (demonstrating, ascribing, partitioning). The activities are a part of some normative practice of modelling, but the relations seem independent of that practice. They are at least conceptually distinct since they can be in principle described without appeal to the practice itself.

A deflationary strategy would recommend replacing both denotation and mapping with functional activities or features of the representational practice as well. I have argued (Suárez 2015) that there are functional replacements for both denotation and mapping, referred to as *denotative* and *inferential function* respectively. The resulting Denotative Function-Demonstration–Inferential Function (or DFDIF) account is more faithful to modelling practice because it relates all its various components directly to a number of salient features of the practice of model building. In addition, as I discuss in the next section, it possesses the additional advantage to deal with fictional entities in science in a natural fashion.

4 The DDI Model and the Role of Fictions in Modelling

The recent modelling literature emphasizes how scientific models can represent fictional or imaginary entities, processes, or phenomena. There is no need to review any of the case studies in detail; their upshot is that any adequate account of

[5] See e.g. Van Fraassen (2008), Chap. 6.

scientific representation must accommodate representations with fictional or imaginary targets. To give just one illustrious example, Maxwell's famous vortex model of the ether is of course a representation; and it is a representation even though the various components, including for that matter the ether itself, have a rather dubious ontological status.[6] Thus fictions are a key testing ground for any account of representation and particularly so for those that presuppose representation is a relation. Thus the requirement of denotation would rule out fictional representation. However, Elgin (2009, pp. 77–78) has emphasised that this requirement can be weakened: "A picture that depicts a unicorn, a map that maps Atlantis, and a graph that charts the increase in phlogiston over time are all representations, although they do not represent anything. To be a representation, a symbol need not itself denote, but it needs to be the sort of symbol that denotes". That is to say, a source may function 'as a representation' without actually denoting its target. It is enough that the source has "denotative function", and this function can be carried out without eventuating in actual denotation. In other words, one crucial difference between denotation and denotative function is that the former is a success term (for it is impossible for it to be true that 'x denotes y' unless y is real) but the latter is not (since 'x has denotative function and its purported denotation is y' may be true even though y is not real but imaginary or fictional). And while the former (denotation) requires the latter (denotative function) the converse is not true: Denotative function does not require successful denotation – not even in the long term or in a hypothetical future.

The comparisons with art are very pertinent and enlightening on this point (see Suárez 1999), which surely explains why they get recurrently used in this regard. A portrait always has denotative function but does not always denote. Velázquez's portrait of Pope Innocent VI both denotes and has denotative function; but it would be a mistake to say of any of the series of canvasses that it inspired Francis Bacon to produce that it also denotes in spite of the obvious fact that they too are portraits. Or, consider the case of Leonardo's Mona Lisa, which notoriously raises historical questions concerning whom exactly it denotes, and how. These questions are logically and historically independent of the uncontroversial fact that the portrait has denotative function. Similarly, Maxwell's models of the ether may not denote anything. We nowadays take them to have no referent, even though Maxwell, like any other nineteenth century physicist – at least at the time that he introduced the vortex model of the ether – was certainly committed to a carrier of electromagnetic waves. Yet, his attitude to both vortexes and particularly idle wheels was more nuanced. He thought of both as useful analogies but not as literal descriptions of the mechanisms underlying electromagnetic phenomena. In spite of all this, the models seem to function undeniably in a representational fashion. More particularly, there is no substantial difference between the methodology employed for both demonstration and application in such 'fictional' models and the

[6]For more case studies see the various essays contained in Suárez (2009); Woods (2010). For a discussion of Maxwell's (1961/2) model, see e.g. Nersessian (2008).

methodology employed in non-fictional models, such as that employed by Galileo. The same patterns and rules of inference seem perfectly to apply in both cases regardless of whether or not denotation obtains. Since denotative function allows us to account for a much larger family of bona fide scientific representations it seems reasonable to substitute denotative function in an appropriately extended version of the DDI account. In other words there is a striking asymmetry between "denotative function" and "denotation" that is best understood perhaps, in comparing the concepts' respective extensions, since the extension of the former strictly includes the extension of the latter (No denoting symbol fails to also be in the set of those symbols that possess denotative function; but the converse is not true).[7]

The nature of denotative function may now be further clarified. For as was noted earlier in the paper, one thing that stands in the way of a deflationary reading of Hughes' original DDI account is the appeal to denotation. A deflationary account of any concept eschews any reference to any substantive relation between that concept and anything else other than the use of the concept, or the norms that inform such use. There can be no explicit or covert appeal in its definition to a relation between the concept and the world – beyond the aspects of the world that constitute or inform use. Thus on a deflationary account, 'representation' is not understood as a relation between representational models, on the one hand, and facts, states, effects, phenomena, etc., on the other hand. It is instead essentially related to features of the use of representations. And this is exactly where the crucial difference between denotation and denotative function has bite. While denotation is a relation between symbols in a language system and their putative referents, denotative function is merely a feature of our use of those symbols. More specifically, a symbol has denotative function if its use within some symbolic practice is in accordance with the typical norms applied to denotative symbols in that practice. In other words, what matters for denotation is whether the putative relation obtains; but what matters for denotative function is independent of whether or not the relation obtains. It depends exclusively on features of our use of symbol systems.[8] Consequently, the revised DFDIF account is deflationary also in the sense of connecting all the essential features of representation to some features of use within representational practice.

The "mapping" relation involved in the original DDI account is susceptible to a similar deflationary strategy (see Suárez 2015, pp. 10–11). The crucial function of this 'mapping' relation is to allow for a transfer of the results of the demonstrations carried out on the source over to the target. Thus in Hughes' Galileo's model

[7]Note that the asymmetry does of course not entail that denotative function is in the end also a success term. Denotative function is a more general term that covers cases of successful denotation and cases of unsuccessful denotation alike. Hence it is not *per se* a success term, even though of one of its subclasses certainly is so.

[8]See Elgin (2009, p. 78) for a similar distinction as applied to what she refers to as 'P-representations' as opposed to 'representations-of-P'. The latter are defined by their relation to a particular kind of things, while the former are, by contrast, defined entirely in terms of features of symbol systems – so belonging in that class is entirely determined by compliance with the norms of use within a practice.

example, the overall area of the triangle is interpreted as the space traversed by the body in its motion over the *t* interval. This is a sort of mapping that thus connects an element in the source system (area in the geometrical figure) with an element in the target system (space traversed by the body in motion in the kinematical system). The point of this mapping in practice is to allow some inferences with respect to the target, and in particular the inference that the time *t* that a body in uniform motion takes to traverse *s* is identical to the time taken by a body uniformly accelerated to a greater speed. Thus the 'mapping' relation's functional role is to constrain the set of inferences about the target that may be performed on the basis of a consideration of the source about the target – i.e. what is technically known as the set of legitimate surrogative inferences.

The deflationary thought is then that this constraint can be stipulated independently of any actual relation between the source and the target. In other words, "taking area to stand for space traversed" sets up a rule of inference with, amongst others, the conclusion that equal areas correspond to equal times travelled. It is still the case that certain claims about the source get transferred over to claims about the target, but note that this transference is achieved without any need for an independently existing actual 'relation' or mapping between the source and target. The transference instead, on this view, maps a set of *claims* about the source over to a set of *claims* about the target. But a mapping between claims of some sort and claims of some other sort does not require any relation between the objects of those claims. In particular, a mapping between claims about A and claims about B does not require that B, or A for that matter, be real entities.

To be sure, the discussion above presupposes the standard metaphysical account of relations, whereby a relation between A and B requires both A and B to be real. It may be possible to weaken this postulate by e.g. requiring existence but not physical reality, or by not requiring existence at all. Thus on some accounts abstract entities may enter into relations, even though they are not concrete physical entities; and on some other accounts relations can obtain even amongst fictional entities that do not exist either as concrete physical entities, abstract entities, or any other way. In either of these cases, the account above regarding claims about A and B requiring no relations would be trivially false, and nothing would be gained in pursuing the deflationary strategy. However, this weakening of the standard metaphysics of relations patently amounts to exactly the same deflationary strategy as applied to the very notion of relation. So in fact the same deflationary strategy is enacted here, but at an earlier stage. Hence it is clear that some deflationary strategy will need to be implemented at some or other stage for the claim above to go through regarding claims in the absence of mapping relations between their objects. Whatever strategy that is, it will see the 'mapping' between source and target as merely an inference generation rule that determines the legitimate move from claims about the source to claims about the target. Talk about 'mapping' then is only genuinely responsive to talk about such inferential rules, and a 'mapping' is acceptable (or not) if the rule that it enacts is correspondingly acceptable (or not). It is in particular not possible to assess the propriety of the mapping independently, as it were – by merely looking into the source and target properties and assessing their similarity or resemblance.

For the critical aspect of the 'mapping' does not lie in any relation between their properties but rather in the generation rule for inferences that it enacts. And while it is possible that the inference generating rules laid down also coincide with a genuine mapping between aspects of a real source and a real target, this mapping is of a piece with the set of generating rules and not independent or prior to it. In particular it need not coincide with any recognizable antecedent similarity or resemblance. Thus in Hughes's example of Galileo's model, we would be at a loss to find any similarities or resemblances between the area of the geometrical figure and space traversed in a certain interval in the kinematical system – until of course the correspondence between area and space is set up, and the set of legitimate surrogative inferences is naturally revealed.

Elsewhere I have referred to this function as the surrogative inference generating function, or inferential function for short (Suárez 2015, p. 11ff.), and I have argued that it should take the place of the "interpretation" third stage in Hughes' original account. The resulting Denotative Function – Demonstration – Inferential Function (DFDIF) account is an extension of the DDI account, suitably weakened to accommodate the representation of fictional entities such as Maxwell's model of the ether. Maxwell's vortex model is genuinely a representation of the ether, even though the ether is nowadays known not to be real. The model represents the ether because it has denotative function and its putative referent is the ether; and because it yields empirical predictions regarding the electromagnetic field when it is so interpreted in the light of the features of the ether. Such denotative and inferential functions are successfully carried out without any successful reference or denotation to the ether. They are carried out because the appropriate norms and rules of inference are enacted in the modelling practice that allows its correct use as a tool in inference. The model is used to all purposes 'as if' it denotes the ether and its elements are interpreted in the light of the features that the ether is assumed to possess. If a model of a fictional entity is functionally indistinguishable from a model of a real entity, then from a deflationary point of view it *is* properly a representation.

5 Conclusions

The DFDIF account here developed has two great virtues. Firstly, it accommodates the representation of fictional entities ubiquitous in scientific practice, which the original DDI model cannot do. And second, it displays the genuine deflationary nature of scientific representation. The role of 'denotation' and 'interpretation' is suitably weakened in this account into their corresponding functional roles. Since a representation can have denotative and inferential functions without actually denoting, the DFDIF account is able to accommodate the representation of fictional entities in science. It also shows representation to be essentially deflationary: the carrying out of the appropriate functions in modelling practice is sufficient for representation. There is in particular no need for a relation to obtain between the

source A and the target B. Of course, the target may be real, and a relation between source and target may obtain, even though it is not necessary for representation. Indeed many representational sources are similar to their targets in some relevant respects. In such cases, the relevant functions may be performed via this relation – but it is important to acknowledge that even in these cases representation is not constituted by the relation. On the account provided here, representation is instead constituted by its denotative, demonstrative and inferential functions in modelling practice.

Acknowledgements This article draws extensively on the discussion of related topics in Suárez (2015), particularly Sects. 3 and 4. I am grateful for comments and suggestions to audiences at the BSPS 2013 conference at Exeter, and the EPSA13 conference in Helsinki. Particular thanks to my co-symposiaists at EPSA13, Sorin Bangu, Tarja Knuuttila, Andrea Loettgers as well as two anonymous referees. Financial support is acknowledged from the Spanish Ministry of Economics and Competitiveness (project FFI2014-57064-P), and from the European Commission (under the Marie Curie programme PEF-GA-2012-329430). Figures 1 and 3 are reprinted from Hughes (1997) with permission from University of Chicago Press.

References

Chang, C., & Kleiser, J. (1990). *Model theory*. Amsterdam: North-Holland.
Elgin, C. (1996). *Considered judgement*. Princeton: Princeton University Press.
Elgin, C. (2009). Exemplification, idealization and understanding. In M. Suárez (Ed.), *Fictions in science: Philosophical essays on modeling and idealization* (pp. 77–90). New York: Routledge.
Goodman, N. (1976). *Languages of art*. Indianapolis/Cambridge: Hackett Publishing Company.
Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science, 64*, S325–S336.
Maxwell, J. C. (1961/2). On physical lines of force. Reprinted In Harman (Ed.) (1990). *The scientific letters and papers of James Clerk Maxwell* (Vol. 1, 2, 3). Cambridge: Cambridge University Press.
Nersessian, N. (2008). *Creating scientific concepts*. Cambridge: MIT Press.
Suárez, M. (1999). Theories, models and representations. In L. Magnani et al. (Eds.), *Model-based reasoning in scientific discovery* (pp. 75–83). Dordrecht: Kluwer Academic Publishers.
Suárez, M. (Ed.). (2009). *Fictions in science: Philosophical essays on modelling and idealization*. London: Routledge.
Suárez, M. (2010). Scientific representation. *Philosophy Compass, 5*(1), 91–101.
Suárez, M. (2015). Deflationary representation, inference and practice. *Studies in History and Philosophy of Science, 49*, 36–47
Van Fraassen, B. (2008). *Scientific representation*. Oxford: Oxford University Press.
Woods, J. (Ed.). (2010). *Fictions and models: New essays*. Munich: Georg Olms Verlag.

Part VIII
Philosophy of the Life Sciences and of Psychology

Non Inferiority Drug Trials and the Trade-offs in RCTs

Cecilia Nardini

1 Introduction

The randomised controlled trial or RCT is currently the gold standard for the scientific evaluation of any newly proposed treatment option. Controlled trials involve learning from a difference – along one or more relevant clinical dimensions – between patients in the experimental group, receiving the new treatment, and patients in the control group. The control group can be on a placebo; in many cases, however, ethical reasons do not permit leaving patients untreated and thus mandate the use of an active control. This typically happens for severe conditions where patients are exposed to a risk of serious and irreversible harm if left untreated. In such cases active-controlled trials or ACT are conducted in place of traditional placebo-controlled trials (PCT).

Randomised controlled trials are especially valued from the scientific point of view because of their features warranting impartiality of the result. For instance, random allocation of participants between the experimental and the control group, and blindfolding of both the participants and the treating physicians, warrant that the interest in promoting a particular treatment cannot influence the outcome of the study. However, some methodologists have raised the concern that ACTs may offer these epistemic warrants to a lesser extent than traditional RCTs with a placebo control (Ellenberg and Temple 2000; Temple and Ellenberg 2000). This issue has been analysed in the philosophy of medicine literature (Anderson 2006;

C. Nardini (✉)
PhD in Foundation and Ethics of the Life Sciences,
University of Milan, Milano, Italy

European Institute of Oncology (IEO), Milano, Italy
e-mail: nardini.folsatec@gmail.com

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_26

Howick 2009); my intention is to contribute to this debate by analysing a particular subspecies of ACT, presenting a specific methodological issue: the case of non-inferiority trials.

Non-inferiority (NI) trials are trials in which, rather than attempting to prove that the experimental treatment over-performs the active control, the aim is that of establishing that the new treatment is not worse by an appreciable amount. This apparently small difference has actually a profound impact upon the way NI trials are designed and analysed. In NI trials, indeed, evaluation of the results depends strongly upon the choice of a non-inferiority margin, i.e. the range of tolerable inferiority of the new treatment face the standard used as the active control. The non-inferiority margin depends on objective data but, as we shall see through discussion, it also depends upon contextual considerations. This appears to be particularly problematic in light of the possibility that arbitrary elements of judgement can enter the otherwise objective and impartial procedure of a clinical trial through this route.

The U.S. Food and Drug Administration and the European Medicines Agency have both issued guidelines for the conduction of NI trials (EMA 2000, 2005; FDA 2000, draft guidance), and there is a vast and comprehensive methodological and medical literature on NI trials (D'Agostino et al. 2003; Djulbegovic and Clarke 2001; Fleming 2008; Head et al. 2012; Jones et al. 1996; Piaggio et al. 2006; Schumi and Wittes 2011). A comprehensive philosophical scrutiny of the NI trial design is, on the other hand, still lacking. This is what I aim to provide in this article.

In the first part of the paper I will introduce non-inferiority trials and describe their methodological peculiarities. In the second part, instead, I will turn to a discussion of the most problematic aspect of NI trials in comparison with conventional (superiority) trials. The aim of my discussion is to show that NI trials do not suffer from a special degree of arbitrariness, as compared to conventional trials. To this aim, I will illustrate the process of sizing a conventional trial, i.e. setting the number of participants. I will argue that contextual elements are important in this, likewise in other phases of any clinical trial; furthermore, I argue that the choice to conduct a NI trial is often dictated by the same kind of contextual considerations, in a negotiation of the stringency of the test against other valuable qualities that is actually common to all forms of clinical trials. NI trials should actually be regarded as part of a continuum of yielding test resolution for the sake of ethical or other contextual considerations.

2 Non-inferiority Trials

The standard superiority design enables investigators to conclude with a certain level of confidence that the new treatment is better than the control. In non-inferiority trials, on the other hand, the objective is not to show that the new treatment has a positive advantage over the control, but rather that it is not worse by a clinically relevant amount.

It is clear that non-inferiority studies are a sub-class of active-controlled trials, since a conclusion of non-inferiority can have a rationale only if the comparison is conducted against an active control. The aim of the NI trial is that of demonstrating a *lack* of substantial difference in action between the new treatment and the control; there can be no clinical or commercial interest in showing that a new compound performs as equally well as a placebo. In other words, not all active-control trials are NI; however, if a trial is a NI, it is necessarily active-control. The rationale for conducting a NI trial rather than a conventional superiority one is generally related to the research question. There are situations where we may be confronted with a new therapeutic option that has secondary advantages with respect to the current standard of care, e.g. it is less expensive, more tolerable or it consists in a less invasive procedure. In such cases a superiority trial, designed to detect an improvement in efficacy, would fail to identify a treatment which is therapeutically roughly equivalent to the standard but which possesses the ancillary benefits mentioned above.

Especially in cancer care, many treatments became established in the past despite their toxicity and harmful side-effects because they provided precious chances of survival. Novel treatments that are proposed nowadays cannot in many case improve substantially on the survival advantage; however, drug development has lead in many cases to less harmful treatments, making an improvement in quality of life possible. As an illustration of this concept one can consider the chemotherapeutic cisplatin in the treatment of testicular teratoma. Until the late 1970s, metastatic testicular teratomas in young men were invariably fatal; the introduction of cisplatin raised cure rates from 10 % to 85 %. On the other hand, cisplatin is highly toxic, with a range of harmful effects that include nerve and kidney damage, vomiting and hear loss (Cullen and Stenning 2004). In such context, it is clear that a new drug providing patients with a better quality of life would be desirable even if its cure rates were comparable to that of cisplatin and not superior. Such are the situations where use of a NI trial in place of a conventional superiority trial should be considered. NI trials are becoming increasingly common, particularly in oncology (Tanaka et al. 2012; Tuma 2007). To date, a few anti-cancer agents have been approved by the US Food and Drug Administration based on results of non-inferiority trials, such as pemetrexed, a second-line treatment of non-small cell lung cancer (Hanna et al. 2004) or capecitabine, an adjuvant treatment in metastatic colon cancer (Twelves et al. 2005). However, the non-inferiority design poses some specific issues, which I now turn to examine from the methodological point of view.

3 Some Issues with Non-inferiority

Superiority and non-inferiority trials alike are based upon the methodology of classical hypothesis test which essentially consists in a statistical procedure for ruling out certain hypotheses in light of experimental data. In the case of a superiority trial, investigators are interested in refuting the null hypothesis that

the two treatments are equally effective – so to conclude that the new treatment is superior. In non-inferiority trials, however, the hypothesis under test is a "null hypothesis" of *worse* performance of the new drug. Thus, if the null hypothesis is refuted, it will be possible to conclude that the new treatment is *non-inferior*. This apparently minor difference between superiority and NI trials leads to major consequences in the appraisal of their results.

The first important detail is the fact that the null hypothesis does include the value zero (no difference between treatments) in the case of conventional superiority trials, while in the case of NI trials it does not. This is a reason why non-inferiority trials may lack *assay sensitivity* (Temple and Ellenberg 2000), i.e. lack the capacity to detect an existing difference between treatments. No trial is completely free from small irregularities in the quality of trial conduct, most often deviations from the protocol that regulates selection and management of participants. These irregularities tend typically to reduce the measured value of the difference in performance between the treatment and the control arm of the trial. In superiority trials, since the null hypothesis includes the zero value of the difference, sloppiness or poor adherence to protocol lead to a *failure* to disprove the null hypothesis and thus produce a negative result for the trial. In the case of NI trials, however, the null hypothesis does not include zero. Thus, everything that blurs the difference between the outcome in the experimental arm and in the control, such as protocol violations, will go in the direction of disproving the null hypothesis and point to the *positive* conclusion that the new treatment is not unacceptably worse than the control. In other words, in a NI setting, irregularities naturally occurring in trial conduct, by hiding existing differences between the experimental treatment and the active control, increase the risk of inappropriately concluding that the new drug is non-inferior to the standard and thus is a viable therapeutic option. The problem is not so much in the fact that this effect is present, but rather that its influence upon the study result cannot be properly determined. As Temple and Ellenberg (2000) remark, "it is difficult in any given [NI trial] to determine the extent to which the ability to show potential treatment differences has been diminished by deficiencies in study design and conduct." (p. 459).[1]

Investigators and regulators are aware of this fact and, when performing a NI trial, they generally put special care in ensuring adherence to protocol. A further contrivance consists in counting only actually treated patients, i.e. relying on the per-protocol (PP) metrics instead of on intention-to-treat (ITT) (D'Agostino et al. 2003). In intention-to-treat analysis, every patient who begins the treatment is considered to be part of the trial and included in the analysis, regardless of whether they follow the protocol to the end of the study or not. ITT is generally simpler than other forms of study design and analysis because it does not require observation of compliance, however it will tend to dilute treatment differences as an effect

[1]Temple and Ellenberg refer, in this quote, to the whole class of active-controlled trials; for the reason discussed above, however, this remark applies particularly well to the case of non-inferiority trials

of counting non-compliant patients and dropouts. On the other hand, per-protocol analysis restricts the comparison of the treatments to those patients who actually completed the entire clinical trial according to the protocol. PP analysis reflects the actual treatment difference to a greater extent than ITT analysis, however, since it involves a form of patient selection successive to randomisation, it may introduce bias in the study (Fleming 2008).

This first issue that has been described implies that greater care and attention should be put in conducting a NI trial versus a superiority RCT; it does not point to any deeply methodological issue arising in one form of trial and not the other. The second point we are turning our attention to, however, does.

As described in the beginning, in NI trials investigators are interested in concluding that the performance of the experimental treatment face that of the control is not worse than an established critical margin. In order to put this hypothesis to test, specification of the margin is necessary, but its value is, in principle, arbitrary. The choice of the non-inferiority margin is a particularly sensitive issue. While in superiority trials the null hypothesis is formulated in terms of the 'neutral' value of no difference, in NI trials the null hypothesis is defined in terms of the non-inferiority margin. Therefore, the value chosen for the margin directly influences the result of the analysis.

The problem is described by Head et al. (2012): "In a non-inferiority trial investigators can choose unreasonably wide margins [. . .] that yield lower sample sizes, and thus improve the trial efficiency, i.e. achieve a positive trial result at a minimized cost". Setting a large non-inferiority margin means allowing for a substantial loss of efficacy for the new treatment with respect to the standard. The negative consequences of setting the bar too low for the new treatment in a single trial are clear enough, but there is a more worrying slippery slope argument attached to it, often mentioned as "biocreep" (D'Agostino et al. 2003): when use of NI trials is sequential – i.e. a treatment that graduated from a NI trial is used as an active control in a subsequent NI trial – there exists a risk of having progressive deterioration of the efficacy of agents that are entering the market. Fleming (2008) raises concerns about "substantial risks of eroding the progress made in benefits delivered by current therapies" (p. 329). The issue is all the more serious given that a pharmaceutical sponsor may have substantial levee in deciding the non-inferiority margin for the test of their product. This fact has led some authors to conclude that non-inferiority trials ought not to be conducted since this form of trial overrides patients' interests (Garattini and Bertelé 2002, 2007; Howick 2009). While this position is contentious (Annoni et al. 2013), it highlights the critical aspect of setting the margin and the substantial ethical import of this operation. Who should decide how the margin is set, and how? Indeed, the draft Guidance on NI trials issued by the US Food and Drug Administration describes the choice of the margin as "the single greatest challenge in the design, conduct, and interpretation of NI trials" (Food and Drug Administration 2000, p. 6).

Typically, the margin is formulated in terms of a percentage of the active control's effect, which represents the maximum amount of the effect of the standard that it would be acceptable to give up. As described in detail by Schumi and Wittes

(2011), the choice of the margin can be deliberative or technical. The deliberative approach, involving a consultation with the stakeholders both on the physicians and on the patients' side, is preferred by the European Medicine Agency, while the technical approach, relying mostly on historical data and pre-clinical data about the treatment effectiveness, is favoured by the US FDA. Thus, a number of methodological solutions exist that can reduce the arbitrariness involved in the choice of the non-inferiority margin. Such arbitrariness is, however, implicit in the definition of non-inferiority in NI trials. Non-inferiority is, strictly speaking, a misnomer; actually, the trial design defines a range of tolerable inferiority within which the new treatment will be considered a viable therapeutic option. As Hung et al. (2005) observe, "Determination of the non-inferiority margin depends on what objective the non-inferiority analysis is intended to achieve" (p. 28). Finally, this observation casts doubts upon the status of NI trials as objective and impartial tests, epistemically on a par with conventional superiority trials.

4 The Process of Sizing Trials

In the previous section I have introduced non-inferiority trials and presented a significant epistemic issue that is unique to this form of study design. Due to the characteristics of the methodology of significance test, the result and interpretation of a clinical trial aimed at showing non-inferiority will depend on the possibly arbitrary choice of a non-inferiority margin. At first glance, this aspects of NI trials singles them out from conventional RCTs as more epistemically suspect, given the non-objective content that may affect the choice of the non-inferiority margin. But is this truly a reason for concern? In the following discussion, I will argue that in NI trials we are in fact trading off stringency of the test for other valuable qualities. What is more, I will contend that this kind of trade-off is actually common to all forms of clinical trials, and that NI trials should actually be regarded as part of a continuum of yielding test resolution for the sake of ethical or other contextual considerations.

The concept of *resolution* is of common use in many scientific contexts. Resolution of an optical device or sensor, for instance, refers to the instrument capability at correctly discriminating two points which are very close to one another. In clinical trials, analogously, we can refer to the resolution of the statistical test that is conducted on the trial results as the ability to detect small differences in effectiveness between treatments. The property that most closely instantiates this concept is the *power* of the test, one of its defining features according to the theory of Neyman and Pearson that is underlying most of the statistical methodologies of use in medical research.

In conventional superiority trials, statistical power expresses the ability of a trial to detect a difference of a given size – or larger – between the treatment and control groups. If the experimental treatment is actually more effective than the control by a certain amount, a trial that is adequately powered will likely conclude superiority;

a trial with low power, instead, is likely to miss the difference and conclude that the experimental treatment is no better than the control. Power is always relative to a certain value of the difference that the test will be able to detect. A trial that is not conclusive in ruling out a difference of 0.2 because of low power will have sufficient power for ruling out a difference of 1.0 instead. A number of features of the study contribute to determining its power; most notably, power is determined by the size of the statistical sample, i.e. by the number of patients that are involved in the trial (Dunn and Clark 2009).

It is intuitive enough that the relationship between sample size and power is a direct one. The higher the number of patients involved in a trial, the more the trial result will be precise and reliable and, consequently, so will be the trial's ability to discern an existing difference. In other words, in the design phase of the trial its resolution can be manipulated through the choice of the sample size; the trial can be made sensitive to smaller and smaller differences between treatments by deciding to enrol more patients. However, unlike for optical devices or for scientific tests in other applications, high resolution is not always a desirable quality in clinical trials. High resolution is what allows investigators to detect small differences between treatments or, in other words, marginal improvements provided by the experimental treatment over the control, and such small improvements are not necessarily of interest for physicians and patients. Furthermore, each participant to a clinical trial comes at a significant ethical and economic cost.

For this reason, it is often the case that resolution is a *de facto* bound variable in clinical trials, as testified by Schumi and Wittes (2011): "Sometimes, investigators [. . .] figure out how much money they can spend. From there, they determine the largest trial that they can run, and justify the [value of the difference] after the fact. This (not exactly a secret) is what investigators often do for superiority trials" (p. 9). It has to be noted that in superiority trials, having a low resolution actually means setting up a tougher test for the treatment under study. If the actual difference in effectiveness is smaller than the value that has been chosen, the trial will likely fail to detect it and it will produce an inconclusive result, possibly leading to the necessity to conduct a new study on a larger sample.

The connection between the foregoing discussion and the issue of the choice of non inferiority margin becomes clear once we consider that among the reasons for conducting a NI trial in place of a superiority trial a capital one is the possibility of reducing the sample size that is needed for achieving a reliable conclusion. In the same paper as mentioned above, Schumi and Wittes determine that, depending on the assumption that is made about the new treatment's effectiveness, it requires four to ten times less patients to demonstrate non inferiority instead of superiority. This in case the new treatment is substantially more effective than the control.

Even more interestingly, conducting a superiority trial upon a treatment that does not provide a substantial advantage over the control will generally require an extremely large sample size, because in this case a high resolution is needed. Non-inferiority trials are not good at discriminating small differences but, in a situation like the one here described, a NI trial can afford a much smaller sample size to reach the conclusion that the new treatment is an acceptable therapeutic option. This may

be of interest in a context where therapeutic options are scarce or improvements badly needed; conceptually, use of a NI trial in a situation like the one here described is equivalent to allowing a small number of less effective agents onboard for the sake of not missing promising options.

5 Clinical Trials: A Microscope Rather than a Yardstick

The foregoing discussion shows two things: the first is that the degree of arbitrariness involved in the choice of the non-inferiority margin does not substantially exceed that involved in the process of sizing conventional trials. In the latter case, the value of the difference that is been sought for, and consequently the resolution of the test, are negotiated according to the ethical and practical limitations imposed by the availability of participants. The second point that emerges in discussion is that the choice to conduct a non-inferiority trial represents just one further step in such negotiation of the resolution of the clinical test within the space created by contextual considerations.

In the case of superiority trials, the need to minimise the number of patients that are exposed to the risks and burdens of trial participation often leads to the practice of setting up a tough test for the experimental treatment, having a considerably high probability of failing it if its performance is less than the demanding value chosen for the difference. On the other hand, in deciding to conduct a non-inferiority trial, investigators are accepting to conduct a test that is more inclusive, i.e. that has higher chances of passing treatments that are just marginally more effective and even treatments that are slightly less effective.

Clearly, it is not reasonable to maintain that one of the two courses of action just highlighted – designing the trial as a severe test or rather as an inclusive one – is right while the other is wrong. One or the other will appear as more adequate, according to the goals of the study and to the therapeutic context in which the research takes place. A superiority study will be more adequate if the interest – of the public, the investigators and the sponsors – is in showing that a new treatment over-performs existing options; if, on the other hand, the objective is to show that the new treatment could prove a viable alternative in a setting where effective options are in scarce supply, or where the ancillary advantages it provides are of interest, then the trial should rightly be designed as a non-inferiority study.

It should be noted that the decision here outlined about which trial design is more appropriate for answering a particular clinical question reflects a contextual evaluation rather than a form of arbitrariness. As Ashcroft (1999) observes in a complex and thorough paper, “medical knowledge has a practical, and not a theoretical, character” and therefore “medical enquiry is [...] inescapably value-laden, even though it is laden with rationally corrigible and objective values” (p. 325). Clinical studies should be considered, as a form of medical inquiry, in light of this consideration. Clinical trials differ from other scientific experiments because

they involve human participants and because the objective of a clinical trial is not that of investigating a scientific truth but rather that of establishing valid treatment options in the context of the actual clinical situation. Thus, it is not only natural but also reasonable that elements of contextual judgement should enter both the conduct and the evaluation of clinical studies, while the same is not expected to happen in experiments in other fields of science.

In the case of NI trials, for instance, such judgements turn out to be particularly important precisely in the choice of the non-inferiority margin, where considerations about how much of the effectiveness of the active control it would be acceptable to give up have a central role. It is for this reason, in fact, that Powers (2008) advocates "[m]ore discussion [. . .] among caregivers about what defines a clinically acceptable loss of effect in various diseases rather than basing the selection of margins on sample size alone" (p. 350).

A further illustration of the same point is provided by the similar case of terminating ongoing trials for apparent benefit (Nardini 2013): the decision about whether to stop an ongoing trials is apparently a purely epistemic problem, that can be solved through recourse to appropriate statistical methods. In reality, though, a number of other elements are often taken into account, such as the credibility of the interim result for the medical community, or concurrent results from other studies. Ultimately, subjective and contextual judgements represent a necessary component in the process of devising and conducting a clinical trial for answering a concrete clinical question within a context where the available knowledge, the existing therapies, the nature of the illness and the situation of the patients both within and outside the trial all have to be taken into account.

The foregoing discussion reveals that the negotiations happening in the design phase of a clinical trial should not be regarded as a threaten to the epistemic reliability of RCTs; rather, they are revealing of the flexibility of this scientific tool. Given the richness of ethical and contextual factors attached to the clinical question that a RCT is bound to answer, a one-size-fits-all approach to designing clinical trials is not possible, but not desirable either. Rather than a fixed yardstick along which treatments are compared, a clinical trial should more appropriately be regarded as a microscope, whose level of detail can be chosen according to the features that we want to investigate by it.

For what concerns non-inferiority trials more particularly, this form of trial design should certainly be approached with attention and a tight level of external supervision, given the special difficulties posed by this form of trial design that were discussed in the first part of the paper. In particular, the definition of the non-inferiority margin is identified as a particularly sensitive issue, requiring thorough discussion in order to ensure that losses in quality of the new treatment are acceptable and controllable. However the discussion contained in the second part of the paper shows that NI trials should not be regarded as a second-order, less epistemically valuable form of RCT; rather, this form of trial design represent an helpful addition to the richness of this scientific instrument.

References

Anderson, J. A. (2006). The ethics and science of placebo-controlled trials: Assay sensitivity and the Duhem–Quine thesis. *Journal of Medicine and Philosophy, 31*(1), 65–81.

Annoni, M., Sanchini, V., & Nardini, C. (2013). The ethics of non-inferiority trials: A consequentialist analysis. *Research Ethics, 9*, 109–120.

Ashcroft, R. (1999). Equipoise, knowledge and ethics in clinical research and practice. *Bioethics, 13*(3/4), 314–326.

Cullen, M., & Stenning, S. (2004). Clinical trials with moving targets: A commentary on a non-inferiority trial in testicular cancer. *The Lancet Oncology, 5*(2), 129–132.

D'Agostino, R. B., Massaro, J. M., & Sullivan, L. M. (2003). Non-inferiority trials: Design concepts and issues–the encounters of academic consultants in statistics. *Statistics in Medicine, 22*(2), 169–186.

Djulbegovic, B., & Clarke, M. (2001). Scientific and ethical issues in equivalence trials. *Journal of the American Medical Association, 285*(9), 1206–1208.

Dunn, O. J., & Clark, V. A. (2009). *Basic statistics. A primer for the biomedical sciences* (4th ed.). Hoboken: Wiley.

Ellenberg, S. S., & Temple, R. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Annals of Internal Medicine, 133*(6), 464–470.

European Medicines Agency (2000). Points to consider on switching between superiority and non-inferiority. Available at: www.ema.europa.eu/pdfs/human/ewp/048299en.pdf. Accessed 6 Dec 2012.

European Medicines Agency (2005). Guideline on non-inferiority margin. Available at: www.ema.europa.eu/pdfs/human/ewp/215899en.pdf. Accessed 6 Dec 2012.

Fleming, T. (2008). Current issues in non-inferiority trials. *Statistics in Medicine, 27*(3), 317–332.

Food and Drug Administration (2000). Guidance for industry: Non-inferiority clinical trials. draft guidance. Available at: www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf. Accessed 6 Dec 2012.

Garattini, S., & Bertelé, V. (2002). Efficacy, safety, and cost of new anticancer drugs. *British Medical Journal, 325*, 269–271.

Garattini, S., & Bertelé, V. (2007). Non-inferiority trials are unethical because they disregard patients' interests. *The Lancet, 370*, 1875–1877.

Hanna, N., Shepherd, F. A., Fossella, F. V., et al. (2004). Randomized phase III trial of pemetrexed versus docetaxel in patients with non–small-cell lung cancer previously treated with chemotherapy. *Journal of Clinical Oncology, 22*(9), 1589–1597.

Head, S., Kaul, S., Bogers, A., & Kappetein, A. (2012). Non-inferiority study design: Lessons to be learned from cardiovascular trials. *European Heart Journal, 33*(11), 1318–1324.

Howick, J. (2009). Questioning the methodologic superiority of 'placebo' over 'active' controlled trials. *The American Journal of Bioethics, 9*(9), 34–48.

Hung, H., Wang, S.-J., & O'Neill, R. (2005). A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal, 47*(1), 28–36.

Jones, B., Jarvis, P., Lewis, J., & Ebbutt, A. (1996). Trials to assess equivalence: The importance of rigorous methods. *British Medical Journal, 313*(7048), 36–39.

Nardini, C. (2013). Monitoring clinical trials: Benefit or bias? *Journal of Theoretical Medicine and Bioethics, 34*, 259–274.

Piaggio, G., Elbourne, D. R., CONSORT Group, et al. (2006). Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. *Journal of the American Medical Association, 295*(10), 1152–1160.

Powers, J. H. (2008). Noninferiority and equivalence trials: Deciphering 'similarity' of medical interventions. *Statistics in Medicine, 27*(3), 343–352.

Schumi, J., & Wittes, J. (2011). Through the looking glass: Understanding non-inferiority. *Trials, 12*, 106.

Tanaka, S., Kinjo, Y., Kataoka, Y., et al. (2012). Statistical issues and recommendations for noninferiority trials in oncology: A systematic review. *Clinical Cancer Research, 18*(7), 1837–1847.

Temple, R., & Ellenberg, S. S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. *Annals of Internal Medicine, 133*(6), 455–463.

Tuma, R. S. (2007). Trend toward noninferiority trials may mean more difficult interpretation of trial results. *Journal of the National Cancer Institute, 99*(23), 1746–1748.

Twelves, C., Wong, A., Nowacki, M. P., et al. (2005). Capecitabine as adjuvant treatment for stage III colon cancer. *New England Journal of Medicine, 352*(26), 2696–2704.

Against Sex and Gender Dualism in Gender-Specific Medicine

Maria Cristina Amoretti and Nicla Vassallo

1 Sex and Gender Dualism

In recent decades, the relevance of sex and gender in influencing health and diseases[1] has been increasingly acknowledged. Meanwhile, a new branch of medicine, generally dubbed as gender-specific medicine, has developed rapidly. Broadly speaking, it aims to investigate when and how sex and gender matter in medicine, by focusing on their relationships with human physiology, pathophysiology, clinical features, and the course of diseases. Sex and gender are now considered central variables in studying, preventing, diagnosing, and treating not only some relatively circumscribed sex- and gender-related diseases, but also human health and diseases in general, and throughout the entire lifespan of individual human beings (Legato and Bilezikian 2010; Oertelt-Prigione and Regitz-Zagrosek 2012; Schenck-Gustafsson 2012).

Despite gender-specific medicine was originally meant to attend to females'/women's health needs (due to the fact that medical research and clinical practice were, and to a certain extent still are, focused on males'/men's bodies and conditions, and thus females/women are either neglected or expected to be amenable to the same diagnostic methods, disease models, and treatments devised for

[1]Even if there are important differences between notions such as disease, disorder, illness, injury, and malady, the conceptual issues raised by them in the present context are similar enough to let us put these distinctions aside.

M.C. Amoretti (✉) • N. Vassallo
Department of Classics, Philosophy, and History, Philosophy Section, University of Genoa, Via Balbi 4, Genoa 16126, Italy
e-mail: cristina.amoretti@unige.it; nicla.vassallo@unige.it

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_27

males/men), this discipline should not be thought of as just female/women-specific medicine; rather, what it aims to study are the relationships of *both* sexes and/or genders with medical research and clinical practice, as well as to improve the health and treatment of *both* males/men and females/women.[2] Some definitions are clear on this issue; Legato and Bilezikian (2010, p. xxi) describe this discipline as the study of:

> how the normal function and the experience of disease differ between men and women. It is as dedicated to the study of unique aspects of men's biology as it is to that of women.

Another study (Oertelt-Prigione and Regitz-Zagrosek 2012, p. 1) similarly asserts that gender-specific medicine is characterised:

> by an unbiased comparison between women and men and the inclusion of gender as a sociocultural process into medical hypotheses. It includes the recognition of biological differences among women and men, i.e., sex differences.

These characterisations of gender-specific medicine, however, also show its implicit acceptance of a dangerous bias, the relevance of which is too often underestimated: the dualistic approach with regard to sex and gender.[3] Specifically, looking at the empirical studies of gender-specific medicine, one can see that they either tacitly or overtly take for granted that it is appropriate to trace a sharp distinction between exactly two sexes (male and female) and exactly two genders (man and woman). Of course, this does not mean that gender-specific medicine aims to defend or preserve any dualistic approach with regard to sex and gender. On the contrary, as far as more than two sexes and genders were recognised, this discipline would hopefully promote the inclusion of the missing ones in medical research and clinical practice. To be clear, in what follows we shall limit ourselves to argue that the assumption of sex and gender dimorphism, which as a matter of fact is currently implicit in medicine and gender-specific medicine, is flawed and far from justified.

Considering first the notion of gender, one may appreciate that it is at once a cultural, historical, psychological, and social concept, which as such may vary greatly from culture to culture and from time to time. Many cultural, sociological, and historical studies have convincingly shown that there are not just two alternative and mutually incompatible genders, man and woman, but instead a plurality of different genders, which range gradually from man to woman (Heyes 2000; Salamon 2010; Stryker and Aizura 2013). The term "transgender" is generally considered as

[2]Even if, from a theoretical point of view, the difference between sex and gender is acknowledged by gender-specific medicine, there is still a tendency to misrepresent and confuse the concepts of sex and gender, which are sometimes even regarded as interchangeable. We discussed this problem arguing that a distinction between the notion of sex (which refers solely to biological, anatomical, physiological, and pathophysiological differentiations) and that of gender (which considers cultural, historical, psychological, and social influences and characteristics) may still be useful in medicine (Amoretti and Vassallo 2013b).

[3]We remarked on this particular bias, and discussed other epistemological problems of gender-specific medicine, in previous works (Amoretti and Vassallo 2012, 2013a, b).

an umbrella term, which aims to group together rather different kinds of people who do not recognise themselves within the traditional binary categories of man and woman.

Moving to the notion of sex, many people claim that it is a binary concept, being a biological, physiological, and pathophysiological category. When discussing species that reproduce sexually, as humans do, biologists seem to have an easy method of distinguishing between males and females: those individuals producing smaller gametes (sperm) are males, while those individuals producing bigger gametes (eggs) are females. Even when acknowledging such a distinction, problems arise when one looks more closely at the sexed bodies (Callahan 2009; Dreger 1998a, b; Fausto-Sterling 1993, 2000). To determine whether a particular body is male or female, there are some sexual traits to consider: genetic sex (46-XY vs. 46-XX), gonadal sex (testes vs. ovaries), gametal sex (sperm vs. egg), genital sex (sperm-related vs. egg-related plumbing parts, including both internal and external genitalia), hormonal sex (more testosterone vs. more œstrogen), and somatic sex (secondary sexual characteristics, such as bodily hair or fat distribution). If an individual is 46-XY, has testes, produces sperm, has sperm-related plumbing parts (such as prostate, seminal vesicles, scrotum, and penis), and produces more testosterone, then he is a male; if an individual is 46-XX, has ovaries, produces eggs, has egg-related plumbing parts (such as fallopian tubes, vagina, uterus, clitoris, and labia), and produces more œstrogen, then she is a female. But what if there is a combination of sperm-related and egg-related plumbing parts? What if the testes are combined with some egg-related plumbing parts or, conversely, the ovaries are combined with some sperm-related plumbing parts? What if an individual has gonads of both types? Or ovotestis, that is, gonads containing both ovarian and testicular tissue? What if a 46-XX individual has testes and/or sperm-related plumbing parts or, conversely, if a 46-XY individual has ovaries and/or egg-related plumbing parts? What if an individual's chromosomes are both 46-XY and 46-XX, or neither of them (45-XO, 47-XXY, 47-XYY, 47-XXX, 48-XXXY, 48-XXYY, 48-XYYY, 48-XXXX)?

Intersexuals—who form approximately 1.7–2 % of the global human population (Blackless et al. 2000; Fausto-Sterling 2000)—show that there is no simple and "natural" sex distinction, as sex traits associated with females and males do not necessarily go together and a particular individual can have a varying mixture of them.[4] (Consider also how transsexuals, who deliberately use hormonal and/or surgical technologies to alter their bodies, indicate the weakness of sex dualism.) How then, should gender-specific medicine, and possibly medicine, deal with intersexuality (and transsexuality)? There exist at least the following three ways: (i) Labelling intersexuals (and transsexuals) as somehow diseased; (ii) Admitting

[4]On November 1st, 2013 Germany became the first country in Europe to allow parents to register intersex babies not as male or female. Australians have had the option of selecting "X" as their sex on passport applications since 2011, New Zealanders since 2012. Similar recognitions of a "third" sex can also be found in Bangladesh, India, Nepal, and Pakistan.

the existence of more than two sexes, where sex (in contrast to gender) is still regarded as a purely biological notion; (iii) Denying that sex is a purely biological, descriptive, and stable concept, as it invariably involves cultural, historical, and social factors, and then endorsing either sex dualism or sex pluralism. In the following section, we will defend the second option while analysing the case of intersexuality.[5]

2 Is Intersexuality a Disease Unto Itself?

Contemporary medicine classifies (most) cases of intersexuality as "disorders of sex development" (DSD), because they do not conform to the dualist scheme that defines an individual's sex as "either male or female" (this label has also been introduced in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders, the DSM-V). However, even though intersexuality is comparatively rare, and though it is true that certain types of intersexuality are associated with some diseases (typically sterility[6] or particular metabolic concerns), we shall argue that there is no compelling reason to label intersexuality as a disease unto itself. To support this claim, we will consider some prominent naturalist conceptions of health and disease: the statistical, the evolutionary, the causal-role, and the biostatistical (BST) account.[7]

According to the statistical account, normal functioning must be established by statistical analysis: normal functioning values fall within a certain range of common variation or distribution. Thus, an organism is regarded as diseased to the extent that it, or some parts of its body, fall outside a certain range of normal limits. Under this definition, intersexuals would probably count as diseased. Still, several other conditions, which are not considered diseases unto themselves, are even rarer than intersexuality: red hair, type 0 blood, or excellent eyesight. In contrast, relatively common conditions, such as dental cavities, are considered diseases. Hence, even if intersexuality would count as a disease according to the statistical account, this very account should be deemed hopelessly flawed, as it fails both as a necessary and as a sufficient condition in identifying disease.

Other more interesting accounts of health and disease refer to the biological notions of fitness and adaptation, in order to define what an evolutionary function is:

[5]We do not survey transsexuality, but many of the following reflections apply to it too.

[6]Even if it has been questioned that sterility is a disease unto itself, for the sake of the argument we won't consider this option.

[7]We do not consider normativist accounts because, as they explicitly refer to what is good/bad, flourishing/harmful for someone (subject, culture, etc.), we feel that it would be easier to accommodate them instead with the idea that intersexuality is not a disease unto itself. Also, we do not want to defend any of the naturalist accounts we consider in this study, as our purpose is merely to show that the claim that intersexuality is not a disease is compatible with (almost) all of them.

broadly speaking, the function of a trait is its contribution to fitness, where fitness is defined as the individual's propensity to survive and/or reproduce.[8] One of the most influential characterisation of evolutionary functions, the ætiological account, was introduced by Wright (1973) and then refined by various scholars in slightly different ways. Wright believes that a function of a trait is a specific effect which, via causal history, can explain the presence or persistence of that trait. In particular, it has been argued that this explanation should be given in terms of natural selection: that the (proper) function of a trait is the effect for which it was selected by natural selection in the past (Millikan 1989; Neander 1991). According to this view, an organism is healthy to the extent that its traits perform those particular effects for which they were selected by natural selection in the past. Conversely, an organism is considered diseased to the extent that some trait fails to produce the effect for which it was selected by natural selection in the past. Similarly, Wakefield (1992) exploits Wright's ætiological account to defend a hybrid (both naturalist and normativist) definition of disease. According to him, a condition is a disease if and only if (a) the condition results from the inability of some internal mechanism to perform that particular effect which is part of the evolutionary explanation of the existence and structure of that very mechanism, and (b) the condition causes some harm or deprivation of benefit to the organism, as judged by cultural or social standards.

Broadly speaking, it seems that intersexuals, being typically sterile, would count as diseased because their sex organs are commonly unable to produce the effects for which they were likely selected in the past, that is, roughly, the ability to make the individual's reproduction possible. Nevertheless, not all people classified as intersex are sterile. For instance, mild hypospadias has typically no consequences on penile function. Many 47-XXX females, as well as 47-XXY and 47-XYY males are fertile. Moreover, there is a difference between claiming that intersexuality is a disease unto itself and saying that there is a disease, sterility, which is often associated with intersexuality. Let us suppose that all individuals with blue eyes are visually impaired; obviously, we would not say that having blue eyes is a disease unto itself. Similarly, intersexuality is not a disease unto itself: even if, for example, an individual who has both testes and ovaries (or has ovotestis) is mostly sterile, having both testes and ovaries (or having ovotestis) is not a disease unto itself; the disease is in being sterile (or having nonfunctional gonads). Another quite common condition associated with intersexuality is congenital adrenal hyperplasia (CAH), which in females is often related to a large clitoris but also to fused labia and other masculine characteristics (while in males it has no relevant consequence for genitalia). The gene called CYP_{21} normally produces a protein that catalyses the conversion of progesterone to cortisol, an effect for which it was reasonably selected by natural

[8] Some scholars have criticised the idea that biological fitness is the common goal of all human life, as humans have multiple and different goals. As we have already seen, it has also been questioned that sterility is a disease unto itself. If these claims were correct, then it would be easier to maintain that intersexuality is not a disease. For the sake of the argument, however, we won't take this possibility into account.

selection; if this gene is absent or blocked, as in CAH, progesterone accumulates. According to the ætiological account, an individual affected by CAH would count as diseased, as he or she has a trait that fails to produce the effect for which it was selected by natural selection in the past. However, again, in this case, having "ambiguous" female genitalia does not count as a disease unto itself. Finally, at least as far as Wakefield's definition is concerned, there is no reason to think that intersexuality invariably causes some harm or deprivation of benefit to the organism that displays it.

An alternative characterisation of evolutionary function is the propensity account; in order to determine a trait's evolutionary function, this approach focuses on the trait's current (not past) contribution to fitness (Bigelow and Pargetter 1987). This means that the function of a trait is what the traits of its type do to contribute to individual fitness at the present time, independently to what they have possibly done in the past. To put it another way, the function of a trait is its present propensity to succeed under selection. Thus, as fitness is understood in terms of survival and reproduction, we may say that an organism is healthy to the extent that its physiological traits keep the organism alive and help it to reproduce. An organism is then considered diseased to the extent that it is not able to survive and/or reproduce through the use of its physiological traits.

Again, it seems that intersexuals, being typically sterile, would count as diseased because their sex organs are unable to help them to reproduce. But we have already seen that not all intersex people are sterile, and that even if we consider sterile intersexuals, the disease is sterility, not intersexuality. Even if most forms of intersexuality do not threaten the individual's survival, still others are associated with life-threatening conditions. For example, in some cases CAH (which in females is often related to "ambiguous" genitalia) is also linked to the lack of a hormone needed for salt metabolism, lack of which leads to death if it is not compensated for with cortisol and other hormones. This, however, does not mean that being intersexual because of "ambiguous" genitalia is a disease unto itself. Moreover, given the propensity account, it is even conceivable to suggest that some types of intersexuality might favour the individual who has them. For instance, males affected by androgen insensitivity syndrome (AIS) have a more or less feminine body, because of the presence of receptors that do not bind strongly to testosterone. AIS comes in three major classes: complete, partial, and mild. In the latter case, the individual shows some feminine features, like body hair, fat, and muscle distribution, and sometimes impaired spermatogenesis. Still, mild AIS might possibly be beneficial in environments where less extremely masculine body types are more adaptable to the environmental conditions. In more general terms, we may say that an organism is regarded as healthy or diseased depending on whether it is successful or unsuccessful in coping with the demands of its current environment. In this respect, there is no reason to assume that intersexuals would be generally unable to cope with their environment because of intersexuality.

As far as physiology is concerned, a characterisation of normal functioning in terms of causal role may be judged preferable. According to Cummins (1975), the function of a trait is that effect which causally contributes to the explanation of

more complex capacities of the organism. Then, if a particular token of a trait of an organism is not able to perform whatever it is that other tokens of this trait do, which contributes to the explanation of more complex capacities of the organism, then the token in question is malfunctional—and the organism diseased.

For example, against the background of an organism's capacity to be able to reproduce, the function of testes/ovaries is to produce sperm/eggs. If our aim is to analyse an organism's complex capacity for reproduction, we can say that if a particular token of the testes/ovaries of an organism is not able to produce sperm/eggs, which clearly contribute to the explanation of the organism's complex capacity for reproduction, then that token is malfunctional. However, according to the causal role account, we may have different research concerns from that of analysing an organism's capacity to be able to reproduce, and have thus decided to focus on other complex capacities of the organism. Against a background of another organism's capacity, the function of testes/ovaries would be different and thus they would probably not count as malfunctional. Generally speaking, this means that whether or not intersexuality is a disease depends largely on our research concerns, that is, on the specific complex capacities we are interested in. And an organism's complex capacity of reproduction is of course not our only option.

Finally, the BST, which is the most prominent naturalistic account of health and disease, claims that a disease is a type of internal state, which is either an impairment of normal functional ability (that is, a reduction of functional abilities below typical efficiency), or a limitation on functional ability, caused by environmental agents. According to Boorse (1977, 1997, 2011), typical efficiency, that is, the statistically typical contribution by a part or process within an organism to its fitness, must be determined in regard to a reference class, that is, an age group of a *sex* of a species. Being "atypical" males or females, it seems that intersexuals would count as diseased under the BST. This conclusion is however mistaken, since it possibly holds only as long as one actually embraces sex dualism, which is exactly what we seek to deny in this study.

If we accept sex pluralism, then the category of sex within the reference class will change accordingly. As typical efficiency must be determined in regard to an age group of a *sex* of a species, so it must obviously be determined in regard to three, four, five, or even more sexes. This means that intersexuals would be compared with their own specific sex group, and as a result, they would no longer count as "atypical" and diseased simply because of their sex.[9] Otherwise, one can still try to preserve sex dualism by acknowledging that sex is not a purely biological, descriptive, and clear-cut concept, as it also involves cultural, historical, psychological, and social factors. In this case, cultural, historical, psychological, and social values would help us to decide whether a particular individual is male or female. This strategy, however, would clash with the strong naturalistic approach

[9]This conclusion might be problematic for the BST, as some diseases generally associated with certain forms of intersexuality could become typical within the reference class selected. This puzzle, however, is not new, as other diseases are typical within a reference class.

of Boorse's BST, according to which health and disease are purely theoretical, descriptive, and non-evaluative concepts.

Besides, and more generally, we also believe that it is important not to confuse the concepts of sex and gender, particularly with regard to medicine. In fact, it could be useful to establish whether relevant differences between healthy individuals can be traced back to sex (biological, physiological, pathophysiological) differences, gender (cultural, historical, psychological, social) differences, or a combination of the two variables, as well as to determine when and how these differences affect not only the clinical features and course of disease (for instance, symptoms or reactions to drugs and surgeries), but also the ways with which health care professionals consult and treat patients.

3 Taking Sex and Gender Pluralism Seriously

In the previous section, we argued that intersexuals should not be deemed as diseased purely because of their intersexuality. Similar reflections may also apply to transsexuals. We additionally claimed that, at least as far as medicine is concerned, it would be beneficial to regard sex as a purely biological concept that, in contrast to gender, does not involve cultural, historical, and social factors. If all the above considerations are sound, then there are good reasons to admit that there are more than two mutually exclusive sexes. Here, two possibilities are in order. On the one hand, one could argue that sex categories are displayed in a continuum, like colour categories, with many variant gradations ranging from male to female. On the other hand, one could say that there are other definite and circumscribed categories, in addition to those of male and female; for instance, it has been argued that there are at least five sexes, and most likely even more. Either way, sex dualism would be denied. We have also suggested that there are convincing reasons to take gender pluralism seriously as well. At least two questions arise from this. Firstly, what potential ethical, epistemological, and medical consequences would sex and gender pluralism have for gender-specific medicine? Secondly, how can gender-specific medicine address sex and gender pluralism from a practical point of view?

Taking sex and gender pluralism seriously would have some important outcomes. To begin, it would be easier to recognise that intersexual, transsexual, and transgender people have their own specific physiology, pathophysiology, and health concerns, which up to now have been mostly overlooked and unaddressed by contemporary medicine and gender-specific medicine, because those disciplines are generally more concerned with finding new methods of "disambiguating" and "normalising", with surgeries and hormonal treatments, the sex of all those individuals who do not conform to the "either male or female" dichotomy. Embracing sex pluralism would help physicians and other health care professionals to move away from the almost exclusive focus on genitalia (gender assignment and genital appearance), and thus to address the other significant medical problems and stigmas that intersexuals, transsexuals, and transgenders typically face. Moreover,

embracing sex and gender pluralism would contribute to the consideration of intersexual, transsexual, and transgender people, and to the inclusion of them in clinical trials, medical research, and treatment.

This would be an ethical, epistemological, and medical improvement. From an ethical point of view, gender-specific medicine would neither run the risk to create or reinforce new conditions of exclusion, marginalisation, segregation, inferiority, new stereotypes and preconceptions, nor to presuppose or support, either explicitly or implicitly, a dominant vs. subordinate hierarchy between male/man and "normal"/"standard" on the one hand, and female/woman and "deviant"/"non-standard" on the other. The absence of such conditions would in turn lead to a reduction of dangerous biases, stigmatisations, and prejudices. Besides, it would be important to no longer label different kinds of intersexuality as disorders of sex development. As we have seen, intersexuality is not a disease unto itself, but simply a variant of nature. Less pathologising alternatives could include "divergences of sex development" (Reis 2007) or "differences of sex development" (Diamond and Beh 2008). From an epistemological point of view, relevant medical evidence would not be precluded or ignored in future, thus encouraging growth and advancement of medical knowledge. Likewise, by incorporating new individual experiences in some clinical trials, medical research, and treatment, it would improve our ability to generalise experimental data and thus to produce more reliable and effective results. From a specifically medical point of view, there would be an improvement of preventative strategies, diagnoses, treatments, and healthcare quality. In particular, it is anticipated that there would be a strong reconsideration of early "normalising" surgeries and forced gender assignments.

To conclude, we would like to briefly sketch some possible ways to address sex and gender pluralism from a practical point of view. To begin, at least for relevant diseases and experimental protocols calling for a large number of individuals, we believe that it should be mandatory to establish clinical trials that include intersexual, transsexual, and transgender people, in order to better evaluate whether or not relevant biological, physiological, and pathophysiological differences can be traced back to sex differences, gender differences, or a combination of the two variables. Moreover, it would be important to promote the studies, analyses, and publications that present disaggregated data with respect to sex and gender, but without the presupposition of sex and gender dualism, on the precise assumption that these variables are in fact medically relevant. This simply means to keep encouraging the kind of studies, analyses, and publications that are already developed by gender-specific medicine, but also pushing this discipline to explicitly recognise sex and gender pluralism and thus champion the inclusion of the missing sexes/genders in medical research and clinical practice. We also advocate that, besides sex and gender, other relevant concepts such as "race", ethnicity, nationality, social class, age, religion, and sexual orientation, should be considered and evaluated by gender-specific medicine, and by medicine in general. Paying attention to all these categories—as feminist epistemologies are by now used to do—does not mean to identify the individuals with their "race", ethnicity, nationality, social class, age, religion, and sexual orientation, nor to promote new biases and stereotypes; on the

contrary, it simply means to regard all these variables as possibly significant from a medical point of view, in the same vein as sex and gender. Of course, there is the problem of establishing which of them have an actual medical relevance, but this is an empirical question that should be solved eventually by medical research and practice. At any rate, as long as biases and stereotypes are set aside, it would be pivotal for health care professionals to consider and treat every patient as a unique individual.[10]

References

Amoretti, M. C., & Vassallo, N. (2012). Women and medicine: Some notes from an epistemological point of view. In J. Hu (Ed.), *2nd international conference on applied social science* (pp. 406–411). Newark: IERI.

Amoretti, M. C., & Vassallo, N. (2013a). Is there any problem with gender-specific medicine? *Verifiche, 42*(1–3), 139–156.

Amoretti, M. C., & Vassallo, N. (2013b). Sex and gender concepts in gender-specific medicine. In G. Lee (Ed.), *Social science and health* (Vol. 19, pp. 221–226). Newark: IERI.

Bigelow, J., & Pargetter, R. (1987). Functions. *The Journal of Philosophy, 84*(4), 181–196.

Blackless, M., Charuvastra, A., Derryck, A., Fausto-Sterling, A., Lauzanne, K., & Ellen, L. (2000). How sexually dimorphic are we? Review and synthesis. *American Journal of Human Biology, 12*, 151–166.

Boorse, C. (1977). Heath as a theoretical concept. *Philosophy of Science, 44*, 542–573.

Boorse, C. (1997). A rebuttal on health. In J. M. Humber & R. F. Almeder (Eds.), *What is disease?* (pp. 1–134). Totowa: Humana Press.

Boorse, C. (2011). Concepts of health and disease. In F. Gifford (Ed.), *Philosophy of medicine* (pp. 13–64). Amsterdam: Elsevier.

Callahan, G. N. (2009). *Between XX and XY: Intersexuality and the myth of two sexes*. Chicago: Chicago Review Press.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy, 72*, 741–764.

Diamond, M., & Beh, H. G. (2008). Changes in the management of children with intersex conditions. *Nature Clinical Practice Endocrinology and Metabolism, 4*(1), 4–5.

Dreger, A. D. (1998a). "Ambigous sex" – or ambivalent medicine? Ethical issues in the treatment of intersexuality. *The Hastings Center Report, 28*, 24–35.

Dreger, A. D. (1998b). *Hermaphrodites and the medical invention of sex*. Cambridge, MA: Harvard University Press.

Fausto-Sterling, A. (1993). The five sexes: Why male and female are not enough. *The Sciences*, *March/April*, 20–24.

Fausto-Sterling, A. (2000). *Sexing the body: Gender politics and the construction of sexuality*. New York: Basic Books.

Heyes, C. J. (2000). *Line drawings: Defining women through feminist practice*. Ithaca: Cornell University Press.

Legato, M. J., & Bilezikian, J. P. (Eds.). (2010). *Principles of gender-specific medicine*. Amsterdam: Elsevier.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science, 56*(2), 288–302.

Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science, 58*, 168–184.

[10]We wish to thank the anonymous reviewers for their helpful comments and objections.

Oertelt-Prigione, S., & Regitz-Zagrosek, V. (Eds.). (2012). *Sex and gender aspects in clinical medicine*. London: Springer.
Reis, E. (2007). Divergence or disorder? *Perspectives in Biology and Medicine, 50*(4), 535–543.
Salamon, G. (2010). *Assuming a body: Transgender and rhetorics of materiality*. New York: Columbia University Press.
Schenck-Gustafsson, K. (2012). *Handbook of clinical gender medicine*. Basel: Karger.
Stryker, S., & Aizura, A. Z. (2013). *The transgender studies reader 2*. New York: Routledge.
Wakefield, J. C. (1992). The concept of mental disorder. On the boundary between biological facts and social values. *American Psychologist, 47*, 373–388.
Wright, L. (1973). Functions. *The Philosophical Review, 82*, 139–168.

Biological Essentialism Concerning the Species Category

Edit Talpsepp

1 Introduction

It has been claimed by a number of authors that pre-Darwinian taxonomic practice was based on essentialist assumptions, according to which the membership of a species is defined by a property (or a set of properties) that all its members share. According to the consensus among philosophers of biology, this essentialism is inconsistent with evolutionary theory for three main reasons: (1) defining species on the basis of particular properties that its members share is inconsistent with the gradualness of evolution; (2) essentialist thinking might make us overlook individual variation between species members and 'population thinking', which are the preconditions for the mechanisms of natural selection to work; (3) essentialist thinking leads to taxonomic monism, which might be inconsistent with the pluralistic classification criteria that we allegedly need in the context of modern evolutionary biology.

Biological essentialism is usually taken to mean the ascription of essences to biological species *taxa*, i.e. particular biological species such as *Felis catus* or *Vulpes vulpes*. Problems (1) and (2) can be solved by dropping shared-nature essentialism concerning species taxa and starting to define these taxa via some modern species concept, most of which are based on the relations between taxon members or taxon members and the environment. However, in order to solve the issues related to problem (3), we have to move up to the level of the species *category* that contains all species taxa. This is because the applicability of a species concept to a particular species taxon does not imply its applicability to the whole species category. Species monism, i.e. the possibility of defining the species category on

E. Talpsepp (✉)
Department of Philosophy, University of Tartu, Ülikooli 18, Tartu 50090, Estonia

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_28

the basis of a single species concept, can be seen as essentialism concerning the species *category*, and questions about this sort of essentialism are different from essentialism concerning species *taxa*.

Similarly to the canonical claim that Darwinism is inconsistent with essentialism concerning species taxa, adopting evolutionary theory is also claimed to contribute to the rejection of essentialism concerning the species category. This paper focuses on the latter, analysing the arguments against species monism and some implications of adopting more than one species concept. One of the main questions to ask is whether it is justified to have several equally legitimate (but possibly conflicting) taxonomies, formed on the basis of different species concepts. Another concomitant topic is analysis of the suggestion that, since different species concepts give rise to species taxa of different causal and ontological structure, the species category as such does not exist. My claim is that *if* we want to hold that species taxa exist, we also have to acknowledge the existence of the species category. Rejecting the existence of the species category on the basis that not all the species taxa share a common property is a symptom of what I call *reverse essentialism*.

2 Biological Essentialism, Species Concepts and Evolutionary Theory

Before continuing, I will briefly refer to the distinction between species taxa and the species category. Species taxa – e.g. *Felis catus* and *Vulpes vulpes* – consist of particular organisms belonging to these species. The species category, on the other hand, consists of all species taxa. One of the central problems in biology – the *species problem*, i.e. the question of how to define and identify species – is concerned with how to define *both* a particular species taxon *and* the whole species category.

As was already said, biological essentialism is usually associated with species taxa. Essentialism concerning species taxa means that these taxa are assumed to have some underlying physical property or properties that all the members of a taxon share and that are causally responsible for other, non-essential properties of taxa members. Biological essentialism that is based on shared material properties clashes with Darwinism, firstly because it leads to sharp boundaries and immutability of species taxa – phenomena inconsistent with gradual evolution of these taxa. Secondly, essentialism ignores the importance of what Mayr (1959) calls 'population thinking'. Population thinking stresses the importance of individual variation, which is a necessary precondition for the forces of natural selection to work, whereas, according to essentialist thinking, essences are 'real' and variation only a deviation.[1]

The third reason for rejecting biological essentialism is that it leads to taxonomic monism, which might be inconsistent with the pluralistic taxonomic practice that we

[1]Mayr mostly uses the notion 'typological thinking' instead of 'essentialist thinking'.

need in the context of evolutionary theory, according to a plethora of philosophers of biology. Indeed, the belief that every taxon has a material essence on the basis of which we could define and identify it would lead to classifying biological organisms in only one supposedly correct way.[2]

Considering the difficulties brought along by shared-nature material essentialism, according to which all species members share a certain physical property, Okasha (2002) has suggested *relational essentialism*, which defines species on the basis of their relational properties. His motivation for doing so is the fact that all main modern species concepts – biological, ecological and phylogenetic – are based on relationships between species members or these members and their environment. If these species concepts refer to the essences of species taxa, then the essences would (respectively) be: being the biggest (potentially or actually) interbreeding population; being a population of organisms that occupies a common ecological niche; being a chunk of a genealogical nexus between two speciation events. The main different species concepts reflect different evolutionary forces (interbreeding, natural selection, common ancestry) that underlie the formation and maintenance of species.

Leaving aside the debates about whether relational essentialism really deserves the name of proper essentialism, we can say that it manages to avoid some of the flaws of shared-nature material essentialism. It is not inconsistent with gradual evolution and it does not ignore the importance of individual variation, and hence we can say that it does not imply the main problems associated with essentialism concerning species *taxa*. However, in order to deal with the issues related to the monism/pluralism of classification criteria and taxonomies, we need to move up to the level of the species *category*, and ask questions about essentialism concerning that level. This is because the applicability of a species concept to a particular species *taxon* does not imply its applicability to the whole species *category*. For instance, claiming that it is the essential property of the taxon *Felis catus* to be the biggest potentially interbreeding population does not mean that it is the essential property of *species as such*, i.e. all members of the species *category*.

According to Ereshefsky, "whereas many authors maintain that evolution renders essentialism concerning species taxa outdated, far fewer are willing to allow that evolution renders essentialism concerning the species category obsolete" (Ereshefsky 2001, p. 129). He associates essentialism concerning the species category with species monism (defining the species category on the basis of a single species concept) and criticizes it on the assumption that biological classification should reflect the multiplicity of evolutionary forces that are causally responsible for the formation and stability of species.

[2]NB: the question concerning the plurality of classification criteria differs from the question concerning the plurality of equally legitimate (but possibly conflicting) taxonomies, but I will come back to this later.

2.1 *Interbreeding Relations as the Essence of the Species Category*

For pre-Darwinian taxonomists like Linnaeus, for Darwin himself and even some biologists after Darwin, the reproductive and/or interbreeding relations between organisms explicitly or implicitly underlay the definition of what were taken to be *species*; hence, we could say that these properties and relations served at least some functions of the essence of the species category. However, there are two main issues with taking interbreeding relations, and the biological species concept, as the sole basis for defining the species category. These issues, which I am going to consider below, illustrate how evolutionary theory explains the need for species pluralism.

2.1.1 The Vagueness Problem

The issue of sharply distinguishing between two species is often complicated by the fact that many morphologically distinct populations have a high rate of hybridization – a phenomenon that was not unfamiliar to even pre-Darwinian biologists and taxonomists. The rise of the idea of the evolution of species allowed the vagueness-related difficulties with identifying species taxa on the basis of their interbreeding relations to be attributed to the *gradualness of speciation*: sometimes it is difficult to distinguish between two highly hybridizing species because they have not reached full reproductive isolation. Also among the complexes that involve vague interbreeding are *ring species*.

Ereshefsky writes: "A ring species consists of a geographic ring of populations such that organisms in contiguous populations can successfully mate, but organisms in populations at distant links in the ring cannot successfully mate." The organisms in distant populations of a ring species have different reproductive mechanisms (Ereshefsky 2010).

Ring species and actively hybridizing species are both good examples for demonstrating that sometimes interbreeding relations and the biological species concept are not sufficient for classifying organisms into species taxa. The biological species concept would have a hard time handling populations that seem to evolve separately and yet are not reproductively isolated. According to Hausdorf (2011), in the face of this situation, accepting the biological species would result in false negatives – species in early stages of speciation are not recognized as species.

2.1.2 Asexually Reproducing Organisms

Many groups of organisms reproduce asexually, and the biological species concept simply does not apply to these organisms. As not many asexually reproducing organisms were known before and even during Darwin's day, their lack of sexual interbreeding was not much of a classification issue at that time. However, the

modern taxonomic principles classify asexually reproducing organisms into species, and here again evolutionary theory and the Modern Synthesis can account for the existence of these species. In the light of evolutionary theory we can understand the claim that sexual reproduction itself is an evolved trait, and, on the geological scale, quite a recent one. As Wilkins (2003a) states, separately evolving lineages existed even before sexual reproduction 'appeared', and some other evolutionary force besides reproductive isolation must have kept them separate. By relating the theory of natural selection with the ideas of genetics, the Modern Synthesis helped to explain how common selection pressures keep evolving lineages cohesive even if the organisms constituting these lineages are not connected via the mechanisms of sexual reproduction.

2.2 The Inability of Ecological and Phylogenetic Species Concepts to Serve as the Definition of the Species Category

If the biological species concept cannot serve as the definition of the whole species category, maybe we could use some other species concept as the definition of the species category – after all, unlike the biological species concept, surely both the ecological and phylogenetic species concepts apply to all groups of organisms, including those that reproduce asexually? In answering this, we have to take into account both the taxonomic activity of practising biologists and philosophical considerations. (Note that we cannot really oppose scientific practice to philosophical and theoretical aspects, as all these are closely interrelated and mutually affective.)

2.2.1 The Ecological Species Concept

While the main mechanism underlying the interbreeding species concepts is gene flow, the main mechanisms supposed to underlie the ecological species concept(s) are common selection mechanisms. According to this concept, a species is a set of organisms exploiting (or adapted to) a single niche. A niche is a particular set of resources: "The differences between the form and behaviour of different species are often related to the differences between the ecological resources that these species exploit" (Ridley 1993, p. 353).

The main problem with the ecological species concept is that it is difficult to define 'ecological niche'. As Grant (1992) points out, ecological differentiation (or ecological isolation) is a universal (but not an exclusive or diagnostic) feature of species in particular – differences in ecological preferences are standard features of geographical and obviously ecological races; and obviously, higher taxa often occupy different adaptive zones. Also, in the case of sexually reproducing organisms, interbreeding relations are usually prioritized over ecological relations. As we can see, the ecological species concept fails to give the *sufficient* criteria for belonging to a species taxon – and hence it also fails to serve as the definition of the species category.

2.2.2 The Phylogenetic Species Concept

Mishler and Brandon (1987) characterize the phylogenetic species concept in the following way: "a species is the least inclusive taxon recognized in a classification, into which organisms are grouped because of degree of monophyly [. . .], that is ranked as a species because it is the smallest "important" lineage deemed worthy of formal recognition, where "important" refers to the action of those processes that are dominant in producing and maintaining lineages in a particular case" (Mishler and Brandon 1987, p. 46). A monophyletic group is a set of organisms that share an ancestor and can be distinguished from other such sets.

The main problem with the phylogenetic species concept is that while many biologists assume species to be significant interbreeding and ecological units, the classifications based on the biological and ecological species concepts quite often conflict with those based on the phylogenetic species concept. As Ereshefsky (2001) says, sometimes stable and good ecological and interbreeding species form paraphyletic taxa that would not be accepted as 'species' on the basis of the phylogenetic species concept. While the phylogenetic species concept(s) are usually applied for classifying fossils and asexually reproducing species, in case of the many organisms of currently existing populations, other species concepts are given prevalence.

Agreeing with Wilkins's (2003b) suggestion that the prescriptive elements of philosophy of science should be based on the actual history of science, we should take the actions of practising biologists into account when discussing the issues related to essentialism concerning the species category, including the monism/pluralism of species concepts. Considering that actual taxonomic practice quite often uses different species concepts in case of different groups of organisms, or combinations of the criteria associated with different species concepts, we can say that we need species pluralism in the context of evolutionary theory. And hence essentialism concerning the species category is wrong.

3 Does Species Pluralism Also Lead to Taxonomic Pluralism?

Species pluralism (i.e. the need for more than one species concept) is handled in several different ways, some of which lead to taxonomic pluralism and some of which do not. I define taxonomic pluralism as accepting several equally justified ways of dividing up biological diversity. The position not leading to taxonomic pluralism is that adopted by Mishler et al. (Mishler and Brandon 1987): according to this view there are several legitimate species approaches, but different approaches apply to different organisms, so no more than one approach is applicable to an organism – there is still only one correct way of classifying organisms. Ereshefsky's species pluralism, on the other hand, leads to taxonomic pluralism. According to him, different species approaches often classify the same organisms into different lineages

(Ereshefsky 2001). He suggests accepting different conflicting taxonomies that are formed on the basis of biological, ecological and phylogenetic species concepts.

Accepting different potentially conflicting taxonomies should be justified if they serve different theoretical purposes that cannot be served by a single classification. The fact that different criteria might lead to different classifications is not a sufficient basis for accepting all of them. We also cannot be too liberal about which theoretical and pragmatic considerations we take as a basis for classifying organisms. For the purposes of communication and information retrieval – the same reasons that underlie Dupre's (1999) suggestion that we adopt only one taxonomy as a *general reference scheme* – I believe it is desirable to have as few biological scientific classifications as possible, and this purpose does not have to conflict with the principle of these classifications being based mainly on theoretical considerations.

3.1 Biotaxonomy vs. Ecotaxonomy

One of the reasons we cannot have both a biotaxonomy and ecotaxonomy was already brought out above: it is difficult to identify species on the basis of the ecological species concept, as it does not specify an 'ecological niche' and is usually applied together with the criteria associated with other species concepts. Another complication is distinguishing between interbreeding and ecological relations as different evolutionary forces underlying the formation and maintenance of different lineages, either conflicting or not. Often, when preferring one species concept over another, such as the biological species concept over the ecological species concept, it is for pragmatic rather than theoretical purposes – interbreeding relations are easier to detect, especially as the ecological niche is not a very well-specified notion. However, the reasons we might consider several possibly conflicting taxonomies must be primarily *theoretical* – species concepts do not only carry the function of *identifying* taxa, but also theoretical content about what species are, what is their role and what are the inferences/generalizations that we can make about them and their members. We need several taxonomies if a single taxonomy, based on the combination of different species concepts, does not satisfy our theoretical requirements.

When justifying the need for criss-crossing taxonomies, Ereshefsky claims that each of the three approaches (biological, ecological and phylogenetic) and the resulting taxonomies reflect important components of evolution – sex, selection, or genealogy: "a biological taxonomy fashioned on only one species approach neglects significant aspects of evolution", and "provides an impoverished picture of life on this planet" (Ereshefsky 2001, pp. 139–140, 143). What I claim is that the problem with distinguishing between ecospecies and biospecies is that the evolutionary forces underlying their 'existence' – interbreeding relations (gene flow) and ecological selection – seem to be interdependent. The interdependence of reproductive and ecological divergence can be illustrated by the phenomenon of character displacement.

The traditional accounts of a speciation scenario have focused on how shifts in resource or habitat use may dictate shifts in reproductive characters. However, the scenario of speciation may also be reversed: "the evolution of reproductive characters stemming from selection to minimize reproductive interference could also cause divergence in traits associated with resource acquisition" (Pfennig and Pfennig 2009). For instance, species that segregate in space or time to avoid reproductive interactions may be exposed to novel, underutilized resources, which in turn lead to shifts in traits associated with resource use. Reproductive and ecological character displacement can promote each other, and in some cases the evolutionary chain of events is unclear (ibid.).

Considering this, I think we cannot strictly distinguish between which evolutionary force – selection for reproductive isolation or ecological resource preferences – is responsible for the formation of species, as the evolutionary paths are interdependent. Hence, it is not always clear which sort of divergence actually designates two species being 'more separate' or more advanced in a speciation event. Grant also claims that phenetic and genetic discontinuity between species is correlated with discontinuity in the distribution of ecological niches, and reproductive isolation operates to fix and maintain this phenetic and genetic discontinuity between species (Grant 1992). Again, we can see that it is not really possible to clearly distinguish between reproductive isolation and the discontinuity between ecological niches as separately accounting for the 'reality' of species.

3.2 The Phylotaxonomy vs. Bio- and Ecotaxonomy

Some authors have claimed that classification based on phylogeny is inconsistent with classification(s) based on the biological and ecological species concepts. Ereshefsky's argument for accepting all three species concepts and the criss-crossing classifications deriving from them is that they all highlight the real set of divisions in the organic world (Ereshefsky 2001, p. 139). My position is a bit more methodological. It seems to me that there are many relations in the organic world that could be considered equally 'real' in addition to interbreeding, ecological and phylogenetic relations. Which of these relations should underlie the taxonomic practice depends on what we want the biological taxonomy or taxonomies to reflect.

There are two properties that a biological taxonomy should ideally have: reflecting the evolutionary 'history' of taxa in terms of the phylogenetic relationships between them; and consisting of taxa that serve as theoretically meaningful units, providing theoretically interesting generalizations about their members. Phylogenetic systematics prioritizes the former, ecological and biological species concepts the latter, though that does not mean that the taxa formed on the basis of phylogenetic relations do not provide theoretically interesting generalizations. If we assume that both the interbreeding-ecological and phylogenetic relations should be reflected in biological classification, then we should consider using different (possibly conflicting) taxonomies indeed. Some authors have suggested the idea

of the PhyloCode, a system of nomenclature alternative to the Linnaean hierarchy. The PhyloCode is a formal set of rules governing phylogenetic nomenclature; it only aims to name clades (monophyletic groups), not paraphyletic or polyphyletic groups. It also does not appeal to the traditional ranks of the Linnaean hierarchy when naming taxa.

If we have two taxonomies that are based on different classification criteria, which units of which taxonomy should be called 'species'? If we want to base our discussions and analysis at least partly on studying the properties of the species taxa as classified by practising biologists, then we should also take into account a wider and practical use of the species notion, not only its definition by cladists. If we want a taxonomy to be exclusively based on phylogeny and consist of nested monophyletic groups, then we might say that this sort of taxonomy consists of clades, not species, as species and clades are different sorts of entities. As long as we stick to the notion 'species' in a biological science, perhaps we can admit that the classification of organisms into species does not perfectly reflect the phylogenetic relations (including evolutionary distance) of taxa.

4 Does Species Pluralism Lead to Rejecting the Existence of the Species Category?

The need for species pluralism has led some philosophers of biology to doubt the existence of the species category (Ereshefsky 2010) or the species *rank* (Mishler 1999), the reason for this being that the species formed on the basis of different species concepts are maintained by different evolutionary forces and hence possess different ontological and causal structure. What it means to ask whether an abstract entity such as the species category exists and which is the answer to the question about existence depends on what we take the species category to be. One way to interpret the claim that the species category does not exist would be taking it to mean that the species category is not an essence-based natural kind in its traditional sense: there are no features that would be characteristic to all and only the taxa belonging to this category. However, the claim about the non-existence of the species category seems to be even stronger than that.

Most authors suggest that species *taxa* do exist, even if the species category does not. For instance, Ereshefsky (1999) suggests distinguishing biological diversity into *ecospecies*, *biospecies* and *phylospecies* while denying the existence of the species category. I find it controversial to claim that species taxa exist *as species* or are even more real than the taxa of other taxonomic levels (the latter position is held by some, but not all, authors) whereas the species category itself is non-existent.[3] No classes/kinds/categories are (more) *real* (than others) outside certain theoretical

[3]It is similar to saying that there are certain entities, existing due to their property X, while this property itself does not exist.

contexts, and species taxa are not real *as species* outside the context stating what it is to be a species. Here I agree with Brigandt (2003) in that that there has to be some sort of general species category: otherwise we could not talk about different *species* concepts.[4] Adopting the species category as a disjunctive category would be a natural move here, but that is the option that Ereshefsky (2010) explicitly rejects.

I find the idea of rejecting the existence of the species category on the basis that it lacks an essential property as dubious as rejecting the existence of species taxa for the same reason – especially as suggested by Ereshefsky, who claims that essentialism concerning the species category is wrong. Rejecting the existence of the species category for the aforementioned reasons might be a symptom of *reverse essentialism*. What I mean by reverse essentialism is the sort of reasoning that is parallel to essentialism: the latter derives the existence of certain properties of category members from the existence of the category; the former derives the non-existence of the category from the lack of certain properties. The variability of the biological world and the gradualness of evolutionary processes make both essentialism and reverse essentialism concerning species taxa and the species category inconsistent with evolutionary theory.

Reverse essentialist assumptions might also be present in the claims according to which, due to the lack of a unifying theoretical principle applying to all species taxa, biological taxonomy should not be theory-based. This sort of claim is made by Dupre (2001), according to whom the units of evolution are far too diverse to serve as units of classification, and only the latter definition should stand for the species category and be denoted by the term 'species'. He suggests abandoning the attempts to base biological taxonomy on evolutionary theory and stick to *taxonomic conservatism* i.e. continue using the same taxonomy as people did before the rise of Darwinism. My suggestion is that even if species taxa are heterogeneous and preserved by different evolutionary forces, they might still behave as approximations of units of evolution. If we want our taxonomy to have any theoretical significance, we should drop the essentialist assumption that heterogeneity of species taxa prevents them from being theoretical entities.

5 Conclusion

Shared-nature material essentialism concerning species is inconsistent with evolutionary theory and is abandoned as the result of adopting it. Some problems of shared-nature essentialism can be avoided by adopting relational essentialism, i.e. defining taxa via the relation-based species concepts. Defining species by one species concept can be seen as essentialism concerning the species *category* and it was probably the case before Darwinism, reproductive relations being the main

[4]If we deny the existence of the category the membership of which is based on 'being a species', why should we talk about different concepts of 'being a species'?.

basis for defining species. However, similarly to essentialism concerning species taxa, essentialism concerning the species category also has its problems.

Evolutionary theory helped to account for the problems of the biological species concepts in two main ways: (1) it revealed that sexual reproduction is itself an evolved trait and that some other evolutionary forces were/are responsible for the formation of asexual species now and before the 'appearance' of sexual reproduction; (2) it revealed that speciation is a gradual process, the final result of which is usually assumed to be full sexual isolation, some lineages being good candidates for being species without having developed reproductive isolation (yet). However, taking into account the actual taxonomic practice and the conceptual issues related to each species concept, we can see that none of these are successful at serving as the definition of the species category and neither ecological nor phylogenetic species concept manages to refer to its essence. We need species pluralism, and essentialism concerning the species category is wrong indeed.

Whether we should also adopt different possibly conflicting taxonomies in addition to species pluralism should depend mainly on theoretical considerations. It is not theoretically justified to have different taxonomies, one based on the biological and another on the ecological species concept, since it is not always possible to distinguish between ecological and interbreeding factors underlying the formation and maintenance of species. However, if we want our taxonomies to reflect both phylogenetic history and interbreeding/ecological relations, we should consider adopting plurality of taxonomies, since the phylogenetic classification might not result in species taxa that are good interbreeding or ecological units and vice versa. However, it might still be that there is only one way of dividing biological organisms into *species*, as some authors would call phylogenetic taxa *clades* instead.

Contrary to some authors, I do not think that anti-essentialism concerning the species category leads to the rejection of the existence of this category, *if* we want to assume the existence of species taxa. Rejecting the existence of the species category on the basis that not all species taxa share a definite property in common might be an indication of *reverse essentialism* that, similarly to biological essentialism, ignores the variability of biological diversity and is inconsistent with evolutionary theory. Reverse essentialism might also underlie the claim that, due to the heterogeneity of species taxa, biological taxonomy should not be theory-based.

References

Brigandt, I. (2003). Species pluralism does not imply species eliminativism. *Philosophy of Science, 70*, 1305–1316.

Dupre, J. (1999). On the impossibility of a monistic account of species. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 3–22). Cambridge (MA)/London: MIT Press.

Dupre, J. (2001). In defence of classification. *Studies in History and Philosophy of Biological and Biomedical Sciences, 32*, 203–219.

Ereshefsky, M. (1999). Species and the Linnaean hierarchy. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 285–305). Cambridge (Massachusetts)/London: MIT Press.

Ereshefsky, M. (2001). *The poverty of the Linnaean hierarchy: A Philosophical Study of Biological Taxonomy (Cambridge Studies in Philosophy and Biology)*. Cambridge, UK: Cambridge University Press.

Ereshefsky, M. (2010). Species. In E. Z. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition). http://plato.stanford.edu/archives/spr2010/entries/species/. Accessed 1 Apr 2014.

Grant, V. (1992). Comments on the ecological species concept. *Taxon, 41*(2), 310–312.

Hausdorf, B. (2011). Progress toward a general species concept. *Evolution, 65*(4), 923–931.

Mayr, E. (1959). Darwin and the evolutionary theory in biology. In B. J. Meggers (Ed.), *Evolution and anthropology: A centennial appraisal* (pp. 1–10). Washington DC: The Anthropological Society of Washington.

Mishler, B. D. (1999). Getting rid of species? In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 307–315). Cambridge (MA)/London: MIT Press.

Mishler, B. D., & Brandon, R. A. (1987). Individuality, pluralism, and the phylogenetic species concept. *Biology and Philosophy, 2*, 397–414.

Okasha, S. (2002). Darwinian metaphysics: Species and the question of essentialism. *Synthese, 131*(2), 191–213.

Pfennig, K. S., & Pfennig, D. W. (2009). Character displacement: Ecological and reproductive responses to a common evolutionary problem. *Quarterly Review of Biology, 84*(3), 253–276.

Ridley, M. (1993). *Evolution*. Oxford: Blackwell Scientific Publications.

Wilkins, J. S. (2003a). How to be a chaste pluralist-realist: The origins of species modes and the synapomorphic species concept. *Biology and Philosophy, 18*(5), 621–638.

Wilkins, J. S. (2003b). *The origins of species concepts: History, characters, modes, and synapomorphies*. Dissertation (PhD). University of Melbourne, Department of History and School of Botany. http://exordio.qfb.umich.mx/archivos%20PDF%20de%20trabajo%20UMSNH/Aphilosofia/species.pdf. Accessed 1 Apr 2014.

Two Concepts of Emotional Expression

Trip Glazer

"Emotional expression" is an unfortunately ambiguous term. Scientists define it in two distinct ways, and this difference in definition leads them to employ the term inconsistently. Often, this confusion is harmless, but sometimes it leads to serious misunderstanding. I aim to disambiguate the two senses of the term and to demonstrate that equivocation has interfered with at least two ongoing empirical projects, namely the debate over the universality of emotional expression and the debate over the evolution of emotional expressions as communicative signals.

1 Disambiguating Terms

When scientists investigate the nature of emotional expression, they typically design experiments using an *operational definition* of the term. That is, they specify some process or test that will enable them to identify and measure emotional expressions as they appear in the context of an experiment. However, because scientists approach the expression of emotion from two different angles, they must employ two separate methods of identifying and measuring these expressions. Distinct empirical contexts lead scientists to adopt distinct operational definitions, which, in turn, leads them to categorize two distinct sets of behaviors as emotional expressions. Rather than attempting to assimilate these definitions, or to subsume one under the other, I suggest that we simply accept their differences and draw a hard line between them. There is not one but *two* concepts of emotional expression currently in use in the sciences. So what are they?

T. Glazer (✉)
Department of Philosophy, Georgetown University, Washington, DC, USA
e-mail: trip.glazer@gmail.com

U. Mäki et al. (eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, European Studies in Philosophy of Science 1,
DOI 10.1007/978-3-319-23015-3_29

One approach to studying emotional expression begins with knowledge of what a subject is feeling, and then investigates which behaviors are correlated with those feelings. For instance, psychologists interested in whether emotional expressions are universal or culturally relative might begin an experiment by putting subjects from different cultural backgrounds into conditions that will elicit the same emotional response (e.g. fear), and then record whether the subjects express this emotion in the same way (e.g. by widening their eyes). Call this basic approach to investigating the nature of emotional expression the "*inside-out approach*" (because we begin with knowledge about what is going on inside of someone—i.e. which emotion she feels—and then we study what occurs on the outside—i.e. how she expresses her emotion in her behavior).

The other approach begins with knowledge of how a subject is behaving, and then investigates whether and how observers characterize this behavior as emotionally infused. For instance, psychologists interested in how ethnicity affects emotion recognition might begin an experiment by posing subjects from various ethnicities into anatomically identical expressions, and then record how observers from different ethnicities classify these expressions. Call this approach the "*outside-in approach*" (because we begin with knowledge about what occurs on the outside—i.e. how someone behaves—and then we study what observers think is going on inside of that person—i.e. what she is feeling).

From the inside-out approach we end up with a concept of emotional expression qua *effect*; from the outside-in perspective we end up with a concept of emotional expression qua *sign*. Importantly, these concepts are not co-extensive; each class has members not included in the other class. Allow me now to say more about each.

1.1 Emotional Expressions Qua Effects

As we have seen, scientists who study emotional expressions using the inside-out approach typically begin with assumptions about what a subject is feeling, and then proceed to investigate how that subject's feelings get expressed. They treat the experience of an emotion as the *independent variable* (i.e. what they manipulate in their experiments), and the resulting expression as the *dependent variable* (i.e. what changes as a result of manipulating the emotion felt). Both Charles Darwin and Paul Ekman utilized this approach in their classic studies of emotional expression. In preparation for his book, *The Expression of the Emotions in Man and Animals*, Darwin sent informal questionnaires to correspondents around the globe, asking them how the indigenous peoples of their regions standardly express their emotions. For instance: "When in good spirits do the eyes sparkle?" (2009: 22–23). Following in Darwin's footsteps, Ekman told stories designed to elicit a particular emotional response to the native people of Papua New Guinea, and then asked them to select the appropriate expression from a series of photographs (2007: 8–11). In both cases, the purpose of the study was to find out whether different individuals would express the same emotion similarly, and whether the same individuals would

express different emotions differently. The felt emotion was the starting point, and the corresponding expression was the end result. Studied from this perspective, "emotional expression" is typically operationalized as follows: *identify subjects who are experiencing specific emotions, and then isolate the observable behaviors that correlate with those emotions. These observable behaviors are, by definition, emotional expressions*.

The inside-out approach has enabled scientists to address the following important questions: (1) Does each emotion have a unique expression, or are there some expressions (e.g. smiles) that correspond to multiple emotions (e.g. Keltner 1997)? (2) Are emotions always expressed (e.g. Ekman and Friesen 1969)? (3) Does culture influence how emotions are expressed (e.g. Friesen 1972)? (4) How do human expressions compare to non-human animal expressions (e.g. Chevalier-Skolnikoff 2006)? All are important questions, and all are suitably investigated through the inside-out approach to emotional expression.

The important point for our purposes, however, is that there are many expressions that are captured by this definition yet which are not captured by the definition of an emotional expression qua sign. Such expressions may be grouped into three general categories.

Underdetermined Effects. An increased heart rate is a common effect of emotional arousal. It is so common, in fact, that it directly results from many distinct emotional states, including both anger and joy (Schachter and Singer 1962). A racing heart can therefore be an effect of anger, but it would not thereby be a sign of anger, because no one who perceives the increased heart rate (by itself) would be able to extract any information about *anger,* in particular. At most, a racing heart is a sign of emotional arousal more generally. Underdetermined effects are expressions that can be caused by several distinct emotions, and hence underdetermine for an observer *which* emotional state actually caused them.

Imperceptible Effects. Blushing is a common effect of embarrassment. For people with light colored skin, blushing is also a common sign of embarrassment (Crozier 2010). If you see someone blush, then you can reliably infer that this person is embarrassed. However, blushing is not a sign when it occurs in people with dark colored skin, since in such cases the blushing cannot be perceived by ordinary observers. Even though these blushes *can* be perceived by scientists with special equipment, the point is that these expressions do not serve as signs for people who do not perceive them, even though they clearly are effects of embarrassment. More generally, unperceived effects are those expressions that are not perceived by certain classes of observers, and hence are not conveying any information to them about their underlying cause.

Unrecognizable Effects. We can "read" the emotional behavior of others because people express their emotions in fairly uniform and predictable ways. Smiles are typical expressions of joy; frowns are typical expressions of anger; scowls are typical expressions of contempt; and so on. But sometimes we express our emotions in completely idiosyncratic ways. Suppose that someone were to express her grief once and only once by wiggling her toes. This wiggling of toes

would be an effect but not a sign of grief, since normal observers would not be in a position to decipher this expression. In order for a behavior to count as a sign for an observer, this observer must be able to interpret this behavior as a sign of grief. Unrecognizable effects are idiosyncratic expressions that cannot be recognized as expressions by normal observers.

A significant consequence of the inside-out approach to defining "emotional expression" is that emotional expressions can be neither voluntary nor insincere. Ekman acknowledges this point explicitly: "I propose that all facial expressions of emotion are involuntary; they are never deliberately made. Note, I say all *facial expressions of emotion*, not all facial movements" (1997: 324). In other words, Ekman insists that a simulated emotional expression is not an emotional expression at all. It is like fool's gold, named after that with which it is often mistaken. Ekman presciently notes that while this consequence may ring counter-intuitive to many, it follows logically from his particular approach to investigating the nature of emotional expression. His goal is to explain an important facet of emotionality by answering the question: what behaviors do emotions instinctually cause? Behaviors that are performed voluntarily or which successfully mimic spontaneous expressions of emotion fall into a blind spot and are not analyzed from the inside-out perspective. The importance of accounting for such expressions arises only when we take another perspective on emotional expression, to which I shall now turn.

1.2 Emotional Expressions Qua Signs

When scientists think about emotional expression from an *outside-in* perspective, by contrast, they typically begin with information about how a subject is behaving, and then proceed to investigate what information observers can glean from this behavior (e.g. about the subject's emotions and motivations). Here, the variables are the reverse as before: the observable expression is the *independent variable* (i.e. what the scientist manipulates), whereas the ascription of emotion is the *dependent variable* (i.e. what changes as a result). This approach is taken by scientists whose focus is on *social communication* rather than on *individual psychology*. Ethologists such as Robert Hinde (1985a, b) and Alan Fridlund (1994) adopt this approach in their studies of emotional expressions. Hinde (1985a) claims, for example, that many emotional expressions are not effects of an organism's occurrent emotional state at all, but are rather strategic attempts to negotiate an interaction with a conspecific. He observes that certain birds' threat displays do not predict the likelihood that they will attack; rather, the function of the display is to elicit a response from another bird, which, in turn, *is* predictive of the first bird's subsequent affective response (cf. Griffiths and Scarantino 2009). For Hinde, as well as for Fridlund and others in this tradition, "emotional expressions" name those behaviors that are used by observers to draw conclusions about a subject's present or future affective state. Whether these behaviors are actually caused by an emotion is

a further question that can be asked of them once we have categorized them as emotional expressions. Studied from this perspective, "emotional expression" is typically operationalized as follows: *identify the behaviors that observers do (or would) categorize as emotionally infused, and then isolate what, if anything, observers can infer about the subject's affective state on the basis of this behavior. These initial behaviors are, by definition, emotional expressions*.

The outside-in approach has enabled scientists to address the following questions: (1) Are emotional expressions signals or cues (e.g. Shariff and Tracy 2011)? (2) What kinds of information do these expressions convey (e.g. Ekman and Friesen 2003; Hinde 1985b)? (3) How do observers interpret emotional expressions—holistically or componentially (e.g. McKelvie 1995)? (4) Are expressions reliable sources of information (e.g. Ekman and Friesen 2003; Barrett 2011)? (5) How does age/gender/race/ethnicity/etc. affect observers' ability to interpret these expressions (e.g. Ebner et al. 2010)? Again, these are all important questions, each of which is suitably studied from the outside-in approach to emotional expression.

As before, the important point for our purposes is that many behaviors fall under this concept that do not fall under the concept of an emotional expression qua effect. And, as before, these behaviors may be grouped into three general categories.

Voluntary Signs. Imagine that you've just received a gift from a loved one. You open it to see that it's what you've always wanted, and you causally express your joy in a standard way, e.g. by grinning. However, suppose that you look up to see that your loved one had turned away just as you opened the present, and didn't see your reaction. You want your loved one to know that you like the gift, so when she looks back at you, you voluntarily repeat the same expression. Notice that this expression is perfectly genuine—you really are happy about the gift—but it is performed voluntarily, and hence is not a direct effect of your emotional state. Voluntary expressions are thus signs without being direct effects of emotional arousal.

Inauthentic Signs. Imagine a similar scenario as before, but, this time, suppose that you don't like the gift you've received. You want your loved one to think that you liked the gift, however, and so you voluntarily contort your face into a grin, just as you did in the previous case, only this time it's inauthentic, in the sense that it's not a genuine expression of your occurrent emotion. Because inauthentic expressions are not the effects of emotional processes, they are signs but not effects.[1]

Overdetermined Signs. Suppose now that you are out on a date, and you want to make a good impression. Your date tells a joke, and although it is only

[1]Whether inauthentic expressions ought to count as emotional expressions is a perennial source of debate in the philosophical literature. Green (2007) insists that inauthentic expressions ought not to count, while Davis (1988) insists that they ought indeed to count. My view on the matter is that intuitions vary precisely because there are two concepts of emotional expression in use. The concept of emotional expression qua effect logically excludes the possibility of inauthentic expression while the concept of emotional expression qua sign plausibly includes this possibility.

funny enough to elicit a brief chuckle, you deliberately intensify this chuckle into belly-busting laughter. In other words, although the laughter itself is an involuntary effect of your amusement, you voluntarily exaggerate the laughter for your date's appreciation. Overdetermined signs are partly determined as direct effects of emotional arousal, and partly determined as effects of voluntary acts. As expressions of the *intensity* of the emotion felt, however, they are signs but not effects.

At this point we can visualize the two concepts of emotional expression as a simple Venn diagram with overlapping circles. Many expressions are *both* effects and signs, but neither circle is completely enclosed within the other. There are some behaviors that are effects but not signs, just as there are other behaviors that are signs but not effects. These are, therefore, two distinct concepts of "emotional expression," which must not be conflated.

2 Muddied Waters

Whenever there are two incompatible ways of defining a single term, the threat of equivocation looms large. Are any prominent scientific accounts of emotional expression guilty of equivocating between these two senses of the term? Unfortunately, some are. I shall highlight two cases in which equivocation has exacerbated or covered up difficulties in ongoing empirical projects.

2.1 The Universality of Emotional Expressions

Scientists generally agree that there are universal facial expressions of emotion in humans (Ekman 2007). However, this claim is not without its detractors. I shall examine two recent challenges to the universality thesis, and demonstrate that they have been motivated by an equivocation in terms. Both begin with experimental results that suggest that the recognition of *emotional expressions qua signs* varies across individuals and cultures, but then draws the conclusion that *emotional expressions qua effects* cannot be universal. However, this inference is faulty.

Let us begin with the claim under attack. For half a century, Paul Ekman has been a leading advocate of the universality of emotional expression in humans. His study of facial expressions of emotion in the native people of Papua New Guinea reestablished the legitimacy of Darwin's universality thesis, and in the time since, he and others have amassed experimental findings that support this thesis (Ekman 2007; Ekman and Friesen 2003). As we have already seen, Ekman is fairly explicit in adopting an inside-out perspective on emotional expression, which leads him to apply the concept of emotional expression qua effect to observable behaviors. His results have gained a sizable following, nearly (but not quite) establishing a scientific consensus on the matter.

Ekman's conclusion has been challenged multiple times, but I will focus on two recent challenges in particular. First, in a paper published in 2012, a team of scientists led by Rachael Jack claimed to have generated experimental results that directly disprove the universality of emotional expression. Using a computer graphics program, Jack et al. (2012) generated three-dimensional simulations of every possible combination of human facial movements. They then presented a random series of these simulations to subjects of both Eastern and Western cultures, who were subsequently asked to identify which emotion (if any) was expressed, as well as the intensity of that emotion. This experiment straightforwardly applies the outside-in approach to emotional expression: the observable behavior is the starting point, and the ascription of emotion is the end result. Jack et al. then used their results to reconstruct the "paradigms" of emotional expressions reflected in both cultures, corresponding to the facial movements most often identified as expressions of particular emotions in each. Were these paradigms identical? Although Jack et al. did find substantial cross-cultural agreement about which simulations expressed which emotions with which intensity, some interesting cultural differences did emerge. The Western paradigm of joy included cheek movements that were absent from the Eastern paradigm, the Western paradigm of surprise included brow movements that were absent from the Eastern paradigm, and the Western paradigm of sadness lacked chin movements that were present in the Eastern paradigm. From this evidence, Jack et al. concluded that emotional expressions are not universal, but are instead culturally specific.[2]

I contend that the disagreement between Jack et al. and Ekman is the result of the former equivocating between the two concepts of emotional expression. The fact that one particular facial configuration serves as a reliable *sign* of sadness in Eastern but not in Western cultures does not demonstrate that Easterners, but not Westerners, *causally* express sadness in this way. It could just as well be that Westerners also contort their faces in this way when sad, but that, for whatever reason, observers have difficulty classifying it as the expression of sadness. Jack et al.'s important research demonstrates that individuals from different cultures look to different parts of the face when deciphering facial expressions, but this result does not carry any implications about how these individuals' emotions are causally expressed. We should expect individuals to differ in their recognition of the expression of emotions in others, and we should not be surprised if different cultures emphasize some facial movements over others in the classification of emotional behavior.

A similar challenge has been raised by Lisa Barrett (2011) in her research. Although Ekman is primarily concerned with investigating emotional expressions from the inside-out approach, he frequently appeals to cross-cultural agreement in the recognition of emotional expressions as evidence for his claim that there are pancultural expressions of emotions. Barrett responds by arguing that cross-cultural

[2] A similar experiment, which equivocates harmlessly, was conducted by Ebner, Riediger, and Lindenberger (2010), investigating how *age* affects the classification of emotional expressions.

agreement persists only as long as experimental subjects are forced to choose between a small number of given emotion terms when classifying expressions. If subjects are permitted to supply their own emotion terms, then agreement goes out the window. So Ekman is wrong, she concludes, to assert that expressions of the basic emotions are truly universal.

Again, this challenge to Ekman's thesis is guilty of equivocation. That observers across different cultures disagree about how to classify particular expressions does not imply that the causal expressions themselves vary across cultures. More generally, we should insist that no results gathered from the outside-in approach bear *directly* on questions raised from the inside-out approach. Such results may bear indirectly, by adding plausibility or by casting doubts on a claim, but we should not attempt to use the outside-in approach to confirm or disconfirm Ekman's claim about the universality of emotional expressions qua effects. To that end, we ought to investigate emotional expressions from the inside-out perspective, as Ekman has.

2.2 The Evolution of Emotional Expressions as Signals

There is another empirical project for which equivocation is an issue, not by creating the appearance of disagreement where none in fact exists, but by obscuring a difficulty that has not been adequately addressed. Since the publication of Darwin's *The Expression of the Emotions in Man and Animals*, ethologists have puzzled over why Darwin did not place more emphasis on the *communicative* functions of emotional expressions. On the one hand, Darwin intended his book to refute the claims of Sir Charles Bell, who held that the human face was intelligently designed to communicate emotions, and thus it makes sense that Darwin would want to downplay those functions as much as possible. On the other hand, Darwin's demonstration that emotional expressions first evolved for non-communicative purposes is sufficient to justify his polemic, and his suggestive hints throughout the book that emotional expressions may have later taken on communicative functions call for an evolutionary explanation (e.g. Darwin 2009: 359). Recent work in ethology has focused extensively on substantiating Darwin's suggestion that many emotional expressions may be *exaptations*, or traits that initially evolved for non-communicative purposes but were later co-opted to serve communicative purposes.

A recent survey article by Azim Shariff and Jessica Tracy (2011) on the functions of emotional expressions expounds the dominant, Darwin-inspired "Two-Stage Model" According to this theory, emotional expressions initially evolved for adaptive purposes unrelated to communication. More specifically, they evolved as cascading patterns of physiological changes that would "promote automatic, adaptive responses to recurrent environmental events that pose fitness challenges" (396). For instance, an animal whose eyes would open widely when afraid would be able to locate potential threats and routes of escape more easily, thereby increasing its chances of survival. In this first stage of their evolution, emotional expressions

could potentially serve as "cues," or as traits that happen to convey information as a by-product, but they are not yet "signals," or traits that evolved specifically for the purpose of conveying this information. Over time, however, these expressions may have become *ritualized*—that is, more exaggerated and pronounced—so that they could take on the function of communicating information to observers. As Shariff and Tracy put it:

> As social interaction became more possible and even vital for many species, the adaptive value of these expressions may have shifted toward communication. As a result, the nonverbal behaviors associated with distinct emotions likely underwent ritualization: a process of change well researched in evolutionary zoology whereby an animal's nonverbal displays become exaggerated, more visible, distinctive, and/or prototypic in order to function as reliable and effective signals . . . For emotion expressions, this shift from cue to signal can be thought of as their second stage of evolution—a paradigmatic example of exaptation, the common evolutionary process whereby a feature that evolved for one reason gradually morphs to serve a secondary adaptive function. (2011: 396)

In sum, the two-stage model proposes that emotional expressions qua effects evolve into emotional expressions qua signs through ritualization, defined specifically as the process whereby a behavior becomes increasingly visible and uniform so that observers may identify it more readily.

This is a promising story about the evolution of emotional expression. However, by equivocating between the two senses of "emotional expression," ethologists have made this story sound easier to tell than it truly is. The problem is that this story tells us only how those emotional expressions qua signs that *genuinely* and *involuntarily* express an emotional state evolved from emotional expressions qua effects. It tells us nothing of how *voluntary* and *inauthentic* expressions could likewise come to serve the function of signaling emotions, since these expressions are *not* repurposed emotional expressions qua effects. In essence, by focusing exclusively on genuine, involuntary expressions, ethologists have gerrymandered a class of emotional expressions qua signs that it is well-poised to explain, and have obscured the complementary class of expressions that is comparatively more difficult to explain. The class does not look gerrymandered, however, because it corresponds neatly to those emotional expressions qua effects that are perceptible from a third-person perspective. Allow me to take a step back and explain this point in greater detail.

In order for something to be a signal—that is, in order for something to represent—the possibility of misrepresentation must be present (Millikan 1984). Emotional expressions qua effects may be able to serve as cues, but they cannot *misrepresent* the organism's emotions, and hence they cannot serve as signals. It follows that we may consider emotional expressions to be signals only if we first widen the concept of emotional expression so as to include those behaviors that potentially *misinform* observers about the emotional state of the organism. These misinforming expressions fall under the concept of emotional expression qua sign, but not the concept of emotional expression qua effect. The evolutionary story about the emergence of emotional expressions qua signs must also include an account of how behaviors that are *not* the direct effects of emotions simultaneously evolved

to have the proper function of signaling emotions to observers. Proponents of the Two-Stage Model overlook or downplay the importance of this part of the story and focus entirely on how emotional expressions qua effects became ritualized to serve as emotional expressions qua signs.

Of course, I do not mean to say that the Two-Stage Model is wrong or misguided. On the contrary, I agree that it is precisely the right kind of story to tell. My point is simply that the story is more complicated than many scientists have acknowledged. In order to explain how emotional expressions qua effects evolved into emotional expressions qua signs, an account of ritualization is not enough; we must additionally give an account of the emergence of voluntary and inauthentic expressions.

If we look back to Konrad Lorenz's (2002) original account of ritualization, however, we find a helpful suggestion as to how this problem might be solved. Shariff and Tracy (2011) use the term "ritualization" to refer only to the aforementioned process of a behavior becoming more pronounced and prototypical in order for it to be more easily perceived by others. However, Lorenz used the term to refer also to a related process, which contemporary scientists sometimes call "instrumentalization." If an organism benefits from other organisms giving uptake to its expressions, then its expressions become "instrumental" as they begin to be elicited by the mere presence of observers who can give it beneficial uptake.[3] Take the example of the animal who widens its eyes when afraid. If this animal benefits from other organisms taking the eye-widening to be an expression of fear, then this animal may begin to widen its eyes whenever it would benefit from this uptake, even when it is not afraid.[4] The behavior, in other words, becomes "instrumentalized," which Lorenz correctly regards as an important aspect of ritualization, and which helps us to fill in the gaps of the Two-Stage Model by explaining the emergence of voluntary and inauthentic emotional expressions qua signs.

I have argued in this paper that scientists define "emotional expression" in two distinct ways, which must not be conflated. I then demonstrated that two ongoing empirical projects have been susceptible to equivocation, thereby calling into question the validity or successes of their findings. My hope is that by using the term "emotional expression" more precisely in the future—reserving the *effect* concept for inside-out approaches and the *sign* concept for outside-in approaches—equivocation can be more easily avoided.

[3]Griffiths and Scarantino (2009) cite evidence for the conclusion that human smiles are more reliably elicited by the presence of expectant observers than by environmental stimuli that arouse feelings of joy.

[4]If expressions are too easy to fake, then they may lose their credibility among audiences. See, for instance, Zahavi and Zahavi (1999).

Bibliography

Barrett, L. (2011). Was Darwin wrong about emotional expressions? *Current Directions in Psychological Science, 20*(6), 400–406.

Chevalier-Skolnikoff, S. (2006). Facial expression of emotion in nonhuman primates. In P. Ekman (Ed.), *Darwin and facial expression: A century of research* (pp. 11–90). Los Altos: Malor Books.

Crozier, R. (2010). The puzzle of blushing. *The Psychologist, 23*(5), 390–393.

Darwin, C. (2009). *The expression of the emotions in man and animal*. New York: Oxford University Press.

Davis, W. (1988). Expression of emotion. *American Philosophical Quarterly, 25*(4), 279–291.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods, 42*(1), 351–362.

Ekman, P. (2007). *Emotions revealed*. New York: Holt Paperbacks.

Ekman, P. (1997). Expression or communication about emotion. In N. L. Segel, G. E. Weisfeld, & C. C. Weisfeld (Eds.), *Uniting psychology and biology* (pp. 315–338). Washington, DC: American Psychological Association.

Ekman, P., & Friesen, W. (1969). Nonverbal leakage and clues to deception. *Psychiatry, 32*, 88–105.

Ekman, P., & Friesen, W. (2003). *Unmasking the face*. Los Altos: Malor Books.

Fridlund, A. (1994). *Human facial expression: An evolutionary view*. San Diego: Academic.

Friesen, W. (1972). *Cultural differences in facial expressions in a social situation: An experimental test of the concept of display rules. Unpublished doctoral dissertation*. San Francisco: University of California.

Green, M. (2007). *Self-Expression*. New York: Oxford University Press.

Griffiths, P., & Scarantino, A. (2009). Emotions in the wild. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 437–453). New York: Cambridge University Press.

Hinde, R. (1985a). Expression and negotiation. In G. Zivin (Ed.), *The development of expressive behavior* (pp. 103–116). New York: Academic.

Hinde, R. (1985b). Was 'the expression of emotions' a misleading phrase? *Animal Behaviour, 33*, 985–992.

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences of the United States of America, 109*(19), 7241–7244.

Keltner, D. (1997). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. In P. Ekman (Ed.), *What the face reveals* (pp. 133–157). New York: Oxford University Press.

Lorenz, K. (2002). *On aggression*. New York: Taylor & Francis.

McKelvie, S. J. (1995). Emotional expression in upside-down faces: evidence for configurational and componential processing. *British Journal of Social Psychology, 34*(3), 325–334.

Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*, 379–399.

Shariff, A. F., & Tracy, J. L. (2011). What are emotional expressions for? *Current Directions in Psychological Science, 20*, 395–399.

Zahavi, A., & Zahavi, A. (1999). *The handicap principle*. New York: Oxford University Press.