2023, 9 (1)

ARGUMENTA

The Journal of the Italian Society for Analytic Philosophy

First published 2023 by the University of Sassari

© 2023 The Authors
Produced and designed for digital publication by the *Argumenta* Staff
All rights reserved. No part of this publication may be reproduced, stored or transmitted in any form or by any means for commercial use without the prior permission in writing from *Argumenta*.

Editor-in-Chief
Massimo Dell'Utri
(University of Sassari)

Associate Editor
Massimiliano Carrara
(University of Padova)

Assistant Editors
Sofia Bonicalzi
(Rome 3 University)
Stefano Caputo
(University of Sassari)
Richard Davies
(University of Bergamo)
Silvia De Toffoli
(IUSS – Pavia)
Filippo Ferrari
(University of Bologna and University of Bonn)
Samuele Iaquinto
(University of Genova)
Federica Liveriero
(University of Pavia)
Antonio Lizzadri
(Università Cattolica, Milan)
Marcello Montibeller
(University of Sassari)
Antonio Negro
(University of Genova)
Giulia Piredda
(IUSS – Pavia), Book Reviews
Pietro Salis
(University of Cagliari)

Editorial Board
Carla Bagnoli (University of Modena and Reggio Emilia)
Monika Betzler (Ludwig Maximilians Universität, München)
Elisabetta Galeotti (University of Piemonte Orientale)
David Macarthur (University of Sydney)
Anna Marmodoro (Durham University and University of Oxford)
Veli Mitova (University of Johannesburg)
Nikolaj J. L. L. Pedersen (Yonsei University)
Sarah Stroud (The University of North Carolina at Chapel Hill)

Argumenta is the official journal of the Italian Society for Analytic Philosophy (SIFA). It was founded in 2014 in response to a common demand for the creation of an Italian journal explicitly devoted to the publication of high quality research in analytic philosophy. From the beginning *Argumenta* was conceived as an international journal, and has benefitted from the cooperation of some of the most distinguished Italian and non-Italian scholars in all areas of analytic philosophy.

Contents

Editorial

It is with particular pleasure that I write my usual Editorial Note this time, and that for at least three reasons.

The first is that the present number of *Argumenta* opens with a Special Issue edited by Mario Alai and Francesco Orilia, dedicated to *Epistemological Issues in the Manifest and the Scientific Images*, as its title makes clear. Since this follows the Special Issue published in the previous number and under the title *Logical and Ontological Issues in the Manifest and the Scientific Images*, it is my firm conviction that, in light of the two Special Issues, the reader can avail herself of one of the most comprehensive and updated analyses of a topic of enduring importance on the contemporary philosophical scene.

The second reason is that the number includes the article *Non-Persistent Truths* written by one of the leading and world-renowned Italian philosophers of language, Andrea Bonomi, an article that for him represents the most significant fruit of a long and important inquiry into the topic of temporalism. In particular, Bonomi provides linguistic reasons to support the hypothesis that evaluations yielding different truth-values at different times are perfectly possible. Bonomi's article inaugurates a new category of articles in *Argumenta*: "Target Articles". These are theoretical arguments written by leading authors in the field, directing attention to stimulating new theoretical ideas. Target articles are selected by the Editorial Board from among the research articles accepted for publication and become the focus for critical discussions; up to three short commentaries will be selected via a call for papers and published in the succeeding issue.

Argumenta 9, 1 (2023): 3-4
ISSN 2465-2334
© 2023 University of Sassari

And the third reason is that this is the first time *Argumenta* hosts a "Critical discussion", thus inaugurating a strand that we hope will foster the broadest possible discussion of themes at the centre of current philosophical interest. Opening the strand is Giuseppe Spolaore, who takes a close look at Patrick Todd's book, *The Open Future: Why Future Contingents Are All False*.

The present number also includes three articles that have already appeared in 'early view' (by Mauro Dorato, Brian Garrett, and Giulia Luvisotto and Johannes Roessler), and that have already made and will continue to make significant contributions to discussion in their respective fields.

The number is then rounded off by the section of Book Reviews. We are proud to offer readers three new thoughtful reviews of as many interesting books.

Finally, I would like to thank all the colleagues who have acted as external referees, the members of the Editorial Board, the Editors of the Special Issue, the Editors of the Book Reviews, and the Assistant Editors. All of them have been very generous with their work, advice, and suggestions. Let me mention in particular the team of assistant editors, who once more proved to be the invaluable aid that they have always been.

As usual, the articles appearing in *Argumenta* are freely accessible and freely downloadable, therefore it only remains to wish you:

Buona lettura!

Massimo Dell'Utri
Editor-in-Chief

Argumenta 9, 1 (2023)
Special Issue

Epistemological Issues in the Manifest and the Scientific Images

Edited by

Mario Alai and Francesco Orilia

The Journal of the Italian Society for Analytic Philosophy

Contents

Epistemological Issues in the Manifest and the Scientific Images: An Introduction

Mario Alai and Francesco Orilia***

** University of Urbino Carlo Bo*

*** University of Macerata*

1. The Manifest Image and the Scientific Image

The national PRIN 2017 project "The Manifest Image and the Scientific Image" of the Universities of Macerata, Florence, Rome 3 and Urbino (prot. 2017ZNWW7F_004) was launched in December 2019 to investigate a serious problem in our understanding of the discoveries of contemporary science, extremely deep and rich in momentous practical consequences, but often also very surprising and even paradoxical *vis a vis* our everyday experience of the world. This puzzle, already addressed by Eddington (1928), Husserl (1936) and Sellars (1962), was described by the latter as a clash between the "scientific image" and the "manifest image" emerging from common sense.

The goals of the project were accordingly fixed as follows: (a) Achieving a better understanding of the manifest image, also by recourse to experimental philosophy. (b) Getting a clearer grasp of the scientific image, especially in three areas: the sustainability of scientific realism concerning properties, relations and unobservables; the nature of time as emerging from current physics; the systems of formal logic introduced to achieve higher consistency than that provided by informal logic. (c) Investigating how the two images must be related from the logical, epistemological and metaphysical point of view if they are to be understood as compatible, in spite of their *prima facie* incompatibility.

Over the last three years these goals have been pursued by the investigators and by a number of collaborators to the project. This resulted, *inter alia*, in a number of papers presented and discussed at the two general conferences of Florence (November 29-30, 2021) and Urbino (June 20-21, 2022). This issue of *Argumenta* collects the investigations conducted from a broadly epistemological point of view, while a previous issue (*Logical and Ontological Issues in the Manifest and the Scientific Images*) collects the articles largely dealing with logical and ontological matters.

Of the papers collected here, five (those by Buonocore and colleagues, Galli, Zorzato, Cevolani-Tambolo, and Savojardo) are devoted to goal (b), viz., better understanding the scientific image and how it can be supported (i.e., with the issue of scientific realism). In addition, Cevolani-Tambolo and Savojardo also deal

Argumenta 9, 1 (2023): 9-25
ISSN 2465-2334

© 2023 Mario Alai and Francesco Orilia
DOI 10.14275/2465-2334/202317.int

with the relations between the two images (goal c). Finally, the article by Angelucci and colleagues contributes both to understanding and supporting the manifest image (goal a) and to clarifying the relations between the two images (goal c).

Three of these papers discuss issues in general philosophy of science (Cevolani-Tambolo, Buonocore and collaborators, and Zorzato), while those by Angelucci and collaborators, Galli, and Savojardo concern the special philosophy of three sciences (respectively, evolutionary game theory, artificial intelligence and the neurosciences). Moreover, both Galli and Zorzato discuss the nature and function of models in science.

2. Buonocore, Margoni, Pero: "Conceiving the Inconceivable: An Assessment of Stanford's New Induction"

In recent years Kyle Stanford (2006) has introduced a new powerful argument against scientific realism, the so-called argument from "*unconceived alternatives*" (UA). He points out that

> we have, throughout the history of scientific inquiry and in virtually every scientific field, repeatedly occupied an epistemic position in which we could conceive of only one or a few theories that were well confirmed by the available evidence, while subsequent inquiry would routinely (if not invariably) reveal further, radically distinct alternatives as well confirmed by the previously available evidence as those we were inclined to accept on the strength of that evidence (2006: 19).

This we know because some of those alternatives were subsequently conceived of and found to be better (more probably or approximately true) than the previously accepted theories. It follows then by a natural induction that even today we fail to conceive theories which are better and more probably true than our own, and therefore that our theories are probably false.

Buonocore, Margoni and Pero explain that Stanford's UA argument draws on two classical antirealist arguments, those from the empirical underdetermination of theories and from the pessimistic meta-induction, although allegedly improving on both. It relies on the idea that different theories can be proposed to account for the same evidence. Stanford's argument, however, is less demanding than the classical underdetermination argument, because it does not require alternatives to be empirically equivalent, but simply empirically equally well confirmed. Therefore, the UA argument does not construct its alternatives "parasitically so as to perfectly mimic the predictive and explanatory achievements of our own theories" (Stanford 2006: 18-19) like the traditional underdetermination argument, but it points at genuine theoretical alternatives that, while unconceived up to a certain time, subsequently were actually adopted. Therefore, the UA argument is based on historical evidence, rather than on philosophical speculation.

One might object, we suppose, that this actually weakens the argument, because the very historical evidence showing that certain better alternatives were ignored at earlier times, also shows that they were recognized a later time, so perhaps we might conclude that in the long run the best alternatives (the more approximately true ones) will be found. If not, at least our theories are becoming better and better confirmed, hence, arguably, more and more approximately true and/or verisimilar.

A reply may come from the other strain in Stanford's argument: it is inductive and pessimistic, but unlike the traditional pessimistic meta-induction, his "new induction" focuses on theorists rather than on theories. In fact, it starts by noticing that at all past times scientists had cognitive (intellectual, psychological or sociological) limits, which prevented them from conceiving some better alternatives; then, it inductively argues that scientists will always have similar limits, so concluding that they will always miss many better alternatives. Again, however, a natural rejoinder is that history shows that those limits can be overcome, and they do not prevent us from continuously progressing toward the truth.

Buonocore, Margoni and Pero argue that Stanford's thesis has various problems, which surface once we ask the question:

(Q) Stanford says that at any time *t* many alternatives were empirically well supported but remained unconceived: but were they conceivable o unconceivable at that time?

By 'conceivable' they mean a theory which one could have conceived given the accepted evidential, theoretical, methodological, or metaphysical presuppositions of the time, but remained unconceived because of the subjective intellectual, psychological, or sociological limits of those scientists.

Unfortunately, Stanford does not offer an explicit answer to (Q), but according to the authors textual evidence suggests that he probably thinks of *conceivable* alternatives, for he writes that scientists

> repeatedly *failed* to conceive of scientifically serious and well-confirmed alternatives to their own proposals. [… Such alternatives] were *scientifically serious* even by the standards of the day despite being unconceived and therefore unconsidered by theorists at the time (Stanford 2006: 60; italics added).

Moreover, this failure persisted even "after we came to embrace substantive evidential, metaphysical, and methodological constraints essentially continuous with those of the present day" (2006: 60).

It might be objected that 'scientifically serious' is different from 'conceivable', and in the above quotation from p. 19, Stanford says that at *t* scientists "could" conceive only *one* alternative, so implying that the others could *not* be conceived. Yet, the authors might reply that for Stanford scientists could not conceive those alternatives because of their own personal or sociological limitations, not because those theories were unconceivable. Besides, the mention of "substantive evidential, metaphysical, and methodological constraints" suggests that Stanford has in mind just what the authors mean by 'conceivability'. Furthermore, they notice that one could not talk of a "failure" in conceiving something if it was, in fact, inconceivable.

Actually, it is not clear that by 'fail' Stanford means unsuccess, rather than simple omission or neglect.[1] Even aside from textual evidence, however, it seems that Stanford *should* be concerned precisely with conceivable theories, if his argument must be distinguished from the old pessimistic induction. In fact, if the reason why a theory remained unconceived were that it was unconceivable, then Stanford's argument would be again an induction over theories, like the old pessimistic induction ("at each past time there were many better but inconceivable

[1] In fact, unlike the Italian verb '*fallire*', the English 'to fail' has both these meanings, the evaluative one and the neutral one.

theories, so this is also happening now"), rather than a "new" induction on the limits of scientists, as it is supposed to be ("at each time *t* scientists proved unable to conceive theories which were conceivable at *t*").

However, Buonocore and friends argue that if actually Stanford's argument applies to conceivable UA, as they suggest, then it hinges on conceivability as an *a-temporal* property of *theories*, because it depends only on the relation of theories with evidence, rather than on the *temporal* limits of the *scientists*; in this sense, it is still an induction on theories, after all, rather than on theoreticians, as Stanford claims.

An even more serious flaw is that, if so, his argument cannot be applied to various remarkable cases of theory change, where clearly the theory that in time would supersede the accepted one was not even conceivable. Thus, the scope of the argument would be seriously limited, it would no longer support the generalized antirealist conclusion that *at any time t* scientists fail to conceive better theories that were serious (i.e., conceivable) alternatives even at *t*. For instance, the authors argue that, contrary to Stanford's claim, in Newton's time the Special Theory of Relativity was unconceivable, for lack of those theoretical, empirical and methodological constraints which turned out to be essential to Einstein's theory.

Therefore, since if Stanford refers to *conceivable* UA, he encounters such problems. However, since his answer to question (Q) is not explicit, Buonocore and collaborators also explore what would follow if Stanford instead referred to *unconceivable* UA (or to both conceivable and unconceivable UA). In that case, they argue, his argument would just be that at any time there are countless possible alternatives (conceivable or unconceivable), among which only one is true, and we will never be able to consider all of them in order to choose the true one. Hence, it would no longer concern a *transient but recurrent* underdetermination, as he says, but it would boil down to the traditional argument from (permanent) underdetermination. Besides, realists could argue that even if a better theory was not conceived at time *t* because the lack of the necessary evidential, methodological, and metaphysical presuppositions made it unconceivable, later on, when those presuppositions become available, it will become conceivable, and so it will probably be conceived.

One might worry that nonetheless the one true theory will escape forever, and certain moderate realists are ready to grant this possibility; however, as noticed above, history shows science is progressive, as the successively conceived alternatives are better and better approximations to the truth, and this will satisfy most current realists.

Summing up, the authors question the novelty of Stanford's argument, for no matter whether his UA are conceivable or unconceivable, his induction actually concerns theories, like the old pessimistic induction, rather than theorists, as he suggests.[2] Moreover, if his UA are conceivable the argument does not apply to some of the most important instances of theory change, where the superseding theory was not conceivable until it was actually conceived. If instead Stanford's UA are unconceivable, then his argument does not really differ from the classical underdetermination argument, and it is effective only against certain implausibly strong versions of realism. Of course, a third answer to question (Q) is possible,

[2] For a further reason why Stanford's "new induction" is not any stronger than the old pessimistic induction, see Alai 2019: §3.

viz., that Stanford doesn't care, as he intends to apply his argument to *both* conceivable *and* unconceivable UA. If so, however, the argument has the drawbacks relating to conceivable UA when it applies to them, and those relating to unconceivable UA when it applies to them.

3. Galli: "Structure Representation of Deep-Learning Models: the case of AlphaFold"

The scientific image of the world is largely drawn by using models. Models are a standard tool of scientists, perhaps their main tool, when theoretical science is concerned. Thus, to understand the scientific image and its relations to the manifest image and to how the world actually is, it is mandatory to understand what models are and how they work, i.e., how they represent. Yet, there are various kinds of models and various ways of understanding the very concept of model. No wonder, then, that this topic is so widely discussed in the philosophy of science, and that two of our papers in this issue are concerned with it.

While Zorzato focuses on a particular kind of theoretical models (the so called "fictional" ones), Galli analyses the models produced by a non-human scientist, the deep-learning neural network system AlphaFold, which has proven so successful in predicting the structure of proteins and in other tasks. Even for him, however, the basic question is still the nature of the representational relation between these models and what they represent, and whether it supports scientific realism.

Preliminarily, Galli presents an interesting taxonomy of kinds of models and of different conceptions of the models' function in science: he distinguishes a *similarity conception*, according to which models represent their target systems by being similar to them; an *inferential conception*, according to which the value of models is mainly pragmatic, consisting in the inferences they can license; and a *structuralist conception*, according to which models represent in virtue of an isomorphism they bear to their targets.

He then discusses Knuuttila's (2021) artifactual view of models, a variant of the inferentialist conception, motivated by the fact that one can model not only real existing systems, but also systems which are merely potential or not yet existing. In the latter case, it would seem, models are better seen as artifacts, i.e., as tools for investigating specific phenomena and answering scientific questions. According to Knuuttila, their function is that of exploring the spaces of possibilities and their success needs not be explained by a representational relation holding between them and a target system (which in fact does not exist in this case). Still, the question of what makes one such model successful or unsuccessful is left open.

Subsequently, Galli explains what AlphaFold is, how it works, and which kinds of models it produces. What a protein can do does not depend only on the sequence of amino acids by which it is composed, but, very importantly, also on the way its string of amino acids folds in space. Therefore, AlphaFold can produce 3-dimensional models of proteins, starting from the mere sequences of amino acids. These can be models of actually existing proteins, but also of merely possible proteins. While in the former case the representation relation is given by an isomorphism obtaining between the model and its real target, in the latter case it is given by the fact that the merely possible protein represented by the model exhibits a certain number of modal properties which in fact characterize actual proteins. Therefore, Galli holds that, contra Knuuttila's, even in this case models bear a

(structural) representation relation to real systems. A robust form of scientific realism is thus implicit in his account.

4. Zorzato: "Fiction and Reality: An Uncanny Relationship"

While Galli analyses the models generated by the AlphaFold neural network, Zorzato focuses on a particular kind of theoretical models (the so called "fictional" ones). Antirealists suggest we discard the scientific *image* on the ground that it does not represent actual non-observable reality, or not correctly anyhow. One reason has always been the use of abstraction and models in science, because models are not exact replicas of their intended targets (the real systems they are meant to model): they are not complete replicas, since they are abstract, leave something out, and they are not correct replicas, since some of their features do not correspond to features of their targets. To this, however, realists reply that, as pointed out by Mary Hesse, all models include positive analogues (features we know to reproduce features of the targets), negative analogues (features we know not to reproduce features of the targets) and neutral analogues (features about which we ignore whether they reproduce features of the targets or not). Thus, we use a model only to the extent that it allows us to offer an accurate picture of the target: in describing, explaining, and predicting, we use the positive analogues, discard the negative analogues, and in advancing research we probe the neutral analogues in order to find out whether they are actually positive or negative, so discovering new features of the target.

The so-called "fictional" models, however, defy this defense of realism. They are models in which not only positive analogues (and, tentatively neutral analogues) are exploited to describe, explain, or predict, but also negative analogues. It would seem to follow, therefore, that the resulting descriptions, explanations, or predictions are false. As an example, Alisa Bokulich (2008) discusses the models produced by scientists for *Rydberg atoms*. These are certain light atoms excited to the point that their outermost electrons are at the threshold of ionization. As a result, their size becomes enormous, approaching the dimensions of minute macroscopic particles. Thus, they can be considered as sitting on the threshold between quantum and classical objects. In fact, the spectral lines emitted by these atoms in strong magnetic fields cannot be explained by current quantum theory; instead, they turn out to be nicely explained by assuming that electrons travel on classical orbits (Main et al. 1986). Furthermore, starting from the experimentally observed spectrum, it proves possible to reconstruct the corresponding orbits as described by the classical theory.

Of course, we know that electrons *do not* travel classical orbits, so this is clearly a negative analogue in the model built to account for Rydberg atoms. According to Bokulich, therefore, we cannot be realist about a model of this kind. Yet, it is successfully used to explain and it even allows some sort of prediction; besides, since the atom approaches the dimension of classical particles, something seems to suggest that there might be some truth to it. Therefore, says Bokulich, neither should the model be interpreted instrumentalistically, as a mere calculation device: what is called for is a "moderate" version of realism.

This compromise, however, has been criticized, first of all because it is not clear what exactly "moderate" realism should be, and how it differs from standard scientific realism; besides, it cannot explain how the model, being fictional, can

represent reality and explain: whatever the model may achieve in this respect, if anything, must be parasitic on the theory.

In response, in her contribution to this issue Lisa Zorzato argues that Bokulich's account of fictional models is largely correct, but it doesn't call for any weakening of realism: the use of such models can be explained by "mainstream" scientific realism just like that of ordinary models. By 'mainstream realism' she understands the position of Psillos (1999), in short, the claim that at least some components of scientific theories can be justifiably believed to be true in the correspondence sense of truth.

In order to appreciate her argument, it should be remembered that authors as diverse as Poincaré (1905), Carnap (1927) and Schlick (1938) insisted that knowledge, in particular scientific knowledge, is structural, and Wittgenstein's *Tractatus* (1921) shows that linguistic representation itself is essentially structural (i.e., we can know only the relations among things or parts of things, hence we can know the intrinsic nature of complex things only to the extent that it is given by the relations of their parts, while we ignore the intrinsic nature of simple things). Therefore, the correspondence which for realists exists between representations and reality is a structural correspondence.

Zorzato does not say whether she agrees with contemporary structural scientific realism that scientific theories can represent only structures or not; but in any case, nobody would question that at least *some* scientific knowledge in the realist sense is structural. Now, she points out that real natural systems can be represented by models at a number of hierarchically ordered levels of abstraction (what she calls "the ladder of abstraction"). More precisely, there can be positive analogues in a model at different abstraction levels: this is to say, there are various more or less abstract structures in a model, which in the successful cases structurally correspond to respectively more or less abstract structures of the target system. Schematizing, the model can have a feature at Level 2 which is false of the target's structure at Level 1, but true of its structure at Level 2. Now, this is enough for "mainstream" realism.

For instance, in the case of Rydberg atoms, Zorzato claims that, while the classical orbits of the model are fictional (since there are no such orbits in reality), they play an explanatory role with respect to the behavior of the electrons, because their structure at a certain level of abstraction corresponds to certain structures of the real atomic spectra. In other words, at the level at which orbits are understood just like those of the planets, the model is false. But at a more abstract level, where only certain selected structures of the orbital behavior are considered, those structures can precisely match certain patterns of the emission spectrum.

A possible concern, here, is that the emission spectrum is an empirical structure; hence, it might be objected, the model simply has an instrumental role, saving the phenomena. However, especially in view of the fact that Rydberg atoms resemble classical particles also in other respects (e.g., their size), it seems quite possible that further analysis identify structural correspondences also at a theoretical level. At any rate, further progress of research in this respect appears to be both desirable and possible.

An important take-home lesson, here, is that when the scientific image is given by models, we must distinguish between the literal picture offered by a model, and its *intended* picture, i.e., the structural one. Clearly, the literal picture is closer to the manifest image (because the model is often drawn from ordinary

empirical knowledge, or at least from consolidated scientific results already incorporated in common sense); however, it is false. Thus, the scientific image can be considered as true only if identified with the intended structural picture offered by the model. Moreover, the latter picture bears a structural resemblance to the empirical data patterns, which are one of the facets of the manifest image.

5. Cevolani and Tambolo: "Empirical Success, Closeness to Evidence, and Approximation to the Truth"

Empirical success is the success of a scientific theory or hypothesis in describing, organizing, explaining, and predicting experience, i.e., in accounting for empirical data, i.e., in entailing true empirical propositions. Henceforth it will be called simply "success". Given its empirical nature, anyway, it can be appreciated in a non-theoretical way, i.e., from the vantage point of the manifest image. Thus, it provides an interface between the scientific and the manifest image: when the latter is used as a benchmark for assessing the validity of the former, as in the debates on scientific realism, success figures as a necessary and most important requirement that hypotheses or theories are called to satisfy.

In fact, Cevolani and Tambolo explain that realists are committed to Laudan's (1981: 32-36) "downward path" (DP) and "upward path" (UP), i.e., respectively, the claim that true or approximately true hypotheses or theories are probably very successful, and that very successful hypotheses or theories are probably at least approximately true.

Scientific antirealists are also interested in success; since they deny that science provides theoretical knowledge, they understand the progress of science simply as the idea that science is growingly successful. Moreover, both realists and antirealists can account for the perduring value and utility of falsified hypotheses by pointing out at their success.

Popper held that we cannot ever know whether a hypothesis is true, but only, sometimes, recognize when it is false. Moreover, it's likely that hypotheses we hold now will be falsified in the future. However, even falsified hypotheses may be more or less "similar" to the truth: hypothesis H1 is more *verisimilar* than hypothesis H2 iff H1 has a true content larger than H2, or a smaller false content, or both. Furthermore, if we find that the subsequent and superseding hypotheses are more verisimilar than the earlier and superseded ones, then we know that there is progress in science (Popper 1963). Popper's idea was subsequently developed by a research tradition in which Oddie (1986), Kuipers (1987), Niiniluoto (1987, 1998), Festa (1982), and lately Cevolani himself have been prominent.

This tradition must face two main problems: first, how do you measure the content of a hypothesis? Intuitively, it can be spelled out as its logical strength, or the number of its consequences. Popper's original definition of verisimilitude had a fatal technical flaw which was exposed by Tichy (1974) and Miller (1974), essentially due to the fact that all propositions have infinite consequences. Niiniluoto, Kuipers, Oddie fixed this (roughly) by considering exclusively the number of the atomic propositions "relevant" to the hypothesis H which are entailed by it, and by relativizing the definition to a language. Cevolani and Tambolo's own definition of verisimilitude is

$$vs(H) = \frac{t}{n} - \frac{f}{n}$$

That is, H's verisimilitude (vs) is measured by the difference between the number of true atomic propositions *t* and of the false atomic propositions *f* entailed by H, both weighted by the total number *n* of atomic propositions of the language. In practice, the larger is the *proportion* of true propositions of the language entailed by H and the smaller is that of false propositions, the more verisimilar H is (see also Cevolani et al. 2011, 2013).

The second problem confronting the verisimilitude tradition is how to estimate how much of H's content is true and how much is false, i.e., the numbers of its true and false consequences, respectively. In fact, H's content (i.e., the *t* true propositions and the *f* false propositions entailed by H) includes: (1) H's empirical consequences which we observed to be (1a) true or (1b) false; (2) H's empirical consequences which we have not been able to observe to be true or false; (3) H's theoretical consequences. Thus, the truth-value of propositions in (1) is known, and propositions in (1a) constitute H's success. Instead, the truth-value of propositions in (2) and in (3) (which are many more than those in (1)) must be estimated, and this can be done first and foremost on the basis of H's success and failures, i.e., of the truth-value of propositions in (1).

In this estimation, therefore, success plays the key role; yet, it is a very difficult and risky extrapolation, since the propositions in (1) are so few in comparison with both the propositions in (2) and in (3), and so different in subject from the propositions in (3). Many realists hold that this task can be aided by considering also the "theoretical (or *nonempirical*) virtues" of H (see Alai 2019: §3.2), but anti-realists contend that we will never have enough reasons to justify the claim that any consequence of H is true (or false), except for those in (1) (e.g., van Fraassen: 1980). This is therefore the main focus of contemporary discussions on realism.

Moreover, Cevolani and Tambolo explain that the idea of success is a vague one, and though there are different ways to explicate it precisely, none is completely satisfactory. Hempel (1948) characterized the success of a hypothesis or theory H as its "systematic power", viz. a measure of the proportion of the content of the available evidence E entailed by H. In other words, E is the set of all the atomic propositions of the language currently known to be true, and systematic power is a function of how many of those propositions H entails. Thus, even falsified hypotheses can be more or less successful: for instance, under this characterization a falsified hypothesis H1 turns out to be more successful of a non-falsified hypothesis H2 if H1 includes a wider proportion of E than H2,[3] which can happen when H1 is more informative than H2. On the other hand, this has the undesirable consequence that if H1 entails H2, then H1 is always at least as successful as H2: this is unacceptable, because, for instance, if H1 is built simply by adding to H2 some false or irrelevant claims, H1 is by definition as successful as H2.

This problem is avoided by Kuipers (2000), according to whom H1 is more successful than H2 iff (a) the confirming instances (the true empirical consequences) of H1 are at least as many as those of H2, (b) the disconfirming instances (the false empirical consequences) of H2 are at least as many as those of H1, and (c) H1 has at least one more confirming instance or one less disconfirming instance than H2. In this way, even if H1 entails H2 it may be less successful than H2, for the false or irrelevant surplus content of H1 with respect to H2 may (and typically will) have some disconfirming instances. Unfortunately, however, when success is so defined it becomes impossible for a falsified hypothesis H1 to be

[3] Since a false hypothesis may have some true consequences.

more successful than a non-falsified one H2, because H1 will have at least one more disconfirming instance than H2.

Cevolani and Tambolo take a clue from Zamora Bonilla's (1992, 1996) notion of "estimated truthlikeness" (i.e., roughly, similarity to the evidence), which is defined by him as directly proportional to the portion of the available evidence E entailed by H and inversely proportional to the "rigor", i.e., informativity, or improbability, of E (where E is the set of all the *m* empirical propositions relevant to H currently known to be true). That notion, however, has the drawback that all falsified hypotheses measure 0. Thus, in the present article, Cevolani and Tambolo define success as "similarity to the evidence *es*", where

$$es(H, E) = \frac{t_E}{m} - \frac{f_E}{m}$$

Here t_E is the number of propositions in E entailed by H (hence, its confirming instances), f_E is the number of propositions in E contradicting H (hence, its disconfirming instances), and m is the number of propositions in E. Therefore, the success of H is given by the difference between the ratio of the confirming instances t_E to the m elements of E, and the ratio of the disconfirming instances f_E to the m elements of E. In a nutshell, a successful hypothesis is one that entails a large proportion of the observations (the elements of E) and contradicts a small proportion of them.

This notion has all the advantages of those of Hempel, Kuipers and Zamora Bonilla, but none of their disadvantages: falsified hypotheses may still be successful, success is still directly proportional to the confirming instances and inversely proportional to the disconfirming instances, but a logically stronger hypothesis is not necessarily as successful as a weaker one. Besides, this notion seems to be precisely what scientific antirealists need to account for scientific progress merely in terms of increasing empirical success, without any realist presuppositions, i.e., without assuming that the theoretical content of hypotheses or theories is even approximately or partly true.

Scientific realists, instead, need a clear notion of success in order to argue that if H is approximately true, then it is very successful (the "downward path", DP) and, more importantly, if H is very successful, then it is probably approximately true (the "upward path", UP). UP, of course, is our best bet to estimate verisimilitude.

However, Cevolani and Tambolo show that, if success is constructed as *similarity to evidence* (*es* above), and verisimilarity as *vs* above, neither DP nor UP can be expected to hold in general. For instance, suppose that

(C1a) E is very poor, consisting of just the two propositions p_1, p_2, and suppose H is highly verisimilar.

Yet, quite possibly,

(C1b) H entails p_1 but contradicts p_2. In this case the above definition entails that the similarity of H to evidence equals ½ - ½ = 0; hence, H is highly verisimilar, but without success: DP fails.

Conversely, suppose that

(C2a) As before, E is very uninformative, e.g., consisting of just one proposition p_1, and

(C2b) H entails p_1 (and hence it is maximally successful) but it is extremely poor, to the point of coinciding with p_1 itself.

In this case, although H is maximally successful, its verisimilitude is very low. For instance, if there are 1,000 atomic propositions in the language, by the above definition of *vs*,

$vs\,(\mathrm{H}) = 1/1{,}000 - 0/1{,}000 = 0.001$

Thus, UP is violated in (C2).

Actually, a case like (C2a) is irrelevant to the current debates on scientific realism, for they concern only the possibility of justifying belief in the truth of theoretical hypotheses, while H here is merely empirical. Moreover, few if any realists believe we can show that any hypothesis is (more or less) *verisimilar* (i.e., that it entails most of the propositions of a language, i.e., that it tells a large part of what there is to know, or of "the whole truth"). They are quite content to argue that a hypothesis is (more or less) *approximately true*, i.e., that it is *largely*, or at least *partly*, true (i.e., that most or at least some of its consequences, both empirical and (especially) theoretical, are true—see Musgrave 2006-2007, Alai 2014b: 279-80), *irrespectively of how informative H is*, i.e., of how many propositions it entails: small is beautiful, if it is true. From this point of view, if H entails just itself, and it is true (or, say, if it entails just a theoretical proposition, itself, and an empirical one, and both are true), H is completely true (it has the maximal approximation to the truth), hence, UP works perfectly for it.

Nonetheless, that *es* and *vs* do not support DP and UP can be shown by different examples. For instance, suppose that

(C3a) As above, E consists of just one proposition p_1, correctly entailed by H, so that H's success is maximal, i.e., $1/1 - 0/1 = 1$; still,

(C3b) The theoretical content of H is completely false, and H accounts for p_1 *by pure luck*, or simply because it has been purposefully imagined, or modified *ad hoc*, in order to accommodate p_1.

In this case, then, H is maximally successful, but neither verisimilar nor even slightly approximately true. Therefore, UP fails.

However, Cevolani and Tambolo's formalization shows what is missing from the notion of similarity to evidence *es* to support DP and UP: the definition of *es* is "completely silent on what E is [while] the precise relationship between H and E […] is obviously crucial to assess the success of H on E".

A few comments may be made on this conclusion. First, it might seem that it simply provides a formal confirmation of the intuitive and even commonplace idea that scarce evidence, even if favorable to a hypothesis, cannot confirm it (as required by UP), and that even highly verisimilar and approximately (i.e., largely) true hypotheses might not be successful (as required by DP) at the very first moment, but only in the long run.

This may be right, but there is much more. In fact, even if the body of the *available* evidence E were very large, it would be typically very small (hence of little statistical relevance) with respect to the infinite body UE of the *unavailable* evidence, which escapes us because it is remote in space or time, or beyond the reach of our senses, instruments, or experiments, etc. (UE may be understood as the set of all the true empirical propositions in the language relevant to H but not included in E, i.e. not confirmed by observation). Therefore, it might happen that

(C4a) H is a highly verisimilar and largely true hypothesis, which would get most of UE right, but

(C4b) H it is not successful, because its relatively few empirical failures happen to concern precisely E, hence DP fails.

Conversely, it might be the case that

(C5a) H is quite successful, getting all of E right. Yet,

(C5b) H has a very large and completely false theoretical content,[4] so that most of its empirical consequences are false: in fact, (unbeknownst to us) it contradicts all of UE. Therefore, H is successful but not even partly true,[5] nor verisimilar, and UP is violated.

All this indicates that the main trouble is the gap between *success*, which is empirical, and *truth*, which in the realism-antirealism debate is pre-eminently theoretical, i.e., unobservable, or at least unobserved. Now, the most promising strategy to bridge this gap seems to be one suggested (again) by Popper. In fact, while he initiated the research on *static* properties of hypotheses, like his *verisimilitude*, Hempel's *systematic power*, Kuiper's *empirical success*, and *similarity to evidence*, he also stressed the need to investigate their *dynamics*. An almost trivial example of how dynamic considerations may help in this respect is this: the counterexamples to DP and UP based on the extreme weakness of E (as in cases C1, C2 and C3) can be ruled out because in the absence of a consistent body of evidence to be accounted for, H would not have been proposed in the first place, since there wouldn't have been any need for it, nor enough empirical guidance to conceive it.

This is not all, however, since counterexamples to DP and UP can be envisaged even for large bodies of evidence, like in cases (C4) and (C5). Now, for instance, in a case like (C5a), how can we understand whether (C5b) also holds, i.e., UP is violated, or not? Well, if E was fully known and H was framed precisely to accommodate E, it is very likely that (C5b) holds (i.e., H is neither approximately true nor verisimilar), so that UP fails: in fact, by the principle of empirical underdetermination, there are countless possible *false* hypotheses and only a true one accounting for E. On the other hand, if E was completely unknown beforehand and genuinely predicted by H, by the "no miracles" argument it is overwhelmingly improbable that H is mostly or completely false (Alai 2014a). Hence, it is utterly unlike that (C5b) holds and UP is violated: on the contrary, UP supports the claims that H is at least partly true and to some extent verisimilar.

Theory dynamics also rescues DP, by ruling out cases like (C4). (C4) is impossible because it would be impossible to conceive an almost completely true hypothesis H which is contradicted by all the available evidence E: scientists work out their hypotheses starting from the available evidence. Besides, even if one were so crazy to imagine a hypothesis H which contradicted all the propositions in E, she would have no clue on how to construct H in such a way that all of its theoretical and unobserved empirical content were true. Therefore, it would be cosmically improbable that, among the countless hypotheses contradicting E, she picked just one that happens to be significantly verisimilar or partly true.

[4] This is quite possible if H was shaped *ad hoc* to accommodate E, just like in (C3b) H was shaped to accommodate p_1. More on this below.

[5] Mind, H has a true content, viz., E itself, but this is not in question in the realism-antirealism debate, which, as explained above, is concerned only with the theoretical and the unobserved empirical content of hypotheses (respectively, the propositions in (3) and (2) above).

Therefore, while Cevolani and Tambolo are right that, in order to vindicate DP and UP, we need to take into account "what E is", we need to consider not only the *quantity* of E (i.e., whether it is small, like in (C1), (C2) and (C3), or large, like in (C4) and (C5)), but also its *quality* (e.g., whether it was predicted or just accommodated), as well as the quality of H (e.g., whether it was just constructed *ad hoc* to accommodate E, or it made (also) some daring novel predictions). In other words, as the "predictivist" research tradition has shown (for an overview see Alai 2014a), what matters is more the quality of success than its quantity: even just one novel prediction can confirm more than many pure retrodictions.

6. Savojardo: "The Representation of Reality in the Intelligent Use of Tools"

Savojardo points out how a conflict between the manifest and the scientific image might emerge from the neurosciences. According to the Embodied Cognition account, cognitive activity does not depend only on the brain, but very importantly also on the action of the body on the mind. In particular, in order to avoid an opposition between motor and cognitive aspects, the abilities related to the use of tools are reduced to the sensory-motor level. This opposition won't go away, however: in fact, while the use of familiar tools requires just the retrieval of manipulative sensorimotor information or skills, when we create and use new tools, or use familiar tools in a new way, we need certain specific conceptual skills and certain purely cognitive inferential functions.

Thus, we are threatened by an irreconcilable separation between a prevailingly practical and sensorimotor knowledge, predominant in common everyday use of familiar tools, and a more abstract and theoretical knowledge, especially in science, where instruments themselves become objects of pure reasoning when they are devised, designed, produced and used in order to investigate the world: the manifest image would then become the reign of embodied and sensorimotor cognition, and scientific image the reign of abstract, theoretical knowledge.

According to Savojardo, however, this cleavage may be avoided by two arguments. The former, mainly relying on Buzzoni 2008, begins by maintaining that an intelligent use of tools is essential both in our everyday activities and in science. Whether we use a stick to move in the dark or a probe to explore space, we always do so "guided by an underlying intention to know the environment in order to intervene on it". In any case, "the use of an instrument […] mediates between our body and reality […] and this presupposes an important link between thought and action, and between cognitive and motor elements of knowledge".

Moreover, both in common knowledge and in science "the theoretical moment and the technical moment […] can be distinguished […] only on the level of reflection". Just like in everyday life "the mind often constructs possible alternative scenarios to real situations" to allow successful interaction with the world, at a more elaborate level scientists use counterfactual reasoning to explore aspects of reality more remote from everyday experience. Thus, "there is no human knowledge that is absolutely non-technical, just as there can be no knowledge that is merely practical-technical, unmediated by concept".

Savojardo's second argument hinges on Polanyi's (1958, 1969) distinctions between explicit and tacit knowledge on the one hand and subsidiary and focal awareness on the other. In this relation, commonsense knowledge might appear to be mainly tacit, and scientific knowledge exclusively explicit. Nonetheless,

tools are essential to both, and in both they can either be used automatically, as sensorimotor prolongations of our body, or be explicitly conceived and planned as means to certain cognitive ends. In either way, however, tools are known, although tacitly in the former and explicitly in the latter.

For instance, if I use a hammer to drive a nail, I explicitly consider the hammer and the nail, i.e., I have focal awareness of them and of their operations; however, I couldn't achieve my goal unless, at the same time, I were perfectly aware, although in a merely *subsidiary* and tacit way, of the hammer's impulses on my palm and fingers (Polanyi 1958: 57). We cannot be focused *at the same time* on the instrument with its goals as a whole, and on its details: for instance, a pianist who shifts his attention to his fingers while playing risks to lose sight of the melody.

Nevertheless, whenever it is needed awareness can shift from subsidiary to focal, and *vice versa*, and there are intermediate degrees between them, and thus between purely implicit and fully explicit knowledge, and this clearly applies to the use of scientific instruments as well. For instance, we notice an analogous difference in approach when electrons are studied to investigate their properties, and when they are "sprayed", i.e., used as instruments, to reveal the existence of quarks with fractional charges (Hacking 1983: Ch. 16), or to study the trajectories of neutrons (Giere 1988: Ch. 5).

Therefore, while no doubt the use of tools in common knowledge is largely driven by implicit corporeal knowledge, whereas in science it is largely driven by explicit and highly sophisticated knowledge, this difference is gradual and reversible in perspective and approach, rather than radical.

7. Angelucci, Fano, Ferretti, Macrelli, Tarozzi: "Does Evolution Favor Accurate Perception?"

When one deals with the problem of reconciling the "two images", one usually takes the image s/he considers as more dubious or questionable and tries to understand whether its truth can be proven starting from the other image, which s/he assumes as true by default, or at least as standing on firmer grounds. For instance, in the debates on scientific realism, the manifest image is taken as basically true, and the question is whether the scientific image can stand up to the same standards. Other debates, however, proceed in the opposite direction: according to Sellars himself, it is the scientific image that must be taken as the benchmark for the manifest image,[6] and for philosophers like Paul Churchland (1981, 1988) and Steve Stich (1983) the progress of scientific psychology is showing that the manifest "folk psychology" is radically mistaken.

The paper "Does Evolution Favor Accurate Perception?" by Angelucci and colleagues is of the latter kind: there is a widespread tendency to draw on evolutionary biology to support the reliability of our sensory perception of physical reality, by claiming that in normal conditions our perceptual representations are largely accurate, since natural selection favors epistemically reliable perceptual systems.

This claim, however, has been rejected by Hoffman and colleagues (2013, 2015), who argued that the perceptual systems of animals are adapted to pursue utility (e.g., food, shelter, safety), rather than objective reality. To this end, they

[6] "In the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not" (Sellars 1963: 173).

imagine organisms (call them 'pragmatists') whose perceptual system can distinguish only what is useful to them from what is not, but ignore other objective differences in the environment; on the other hand, imagine organisms (call them 'realists') which can perceptually distinguish a wider range of properties and distinctions. For instance, certain blue things and certain green things may appear to pragmatists of one and the same color (say, grey), since they are all useful, while certain other blue things and green things may again appear to them as sharing one color but a different one (say, brown), because they are not useful (or even harmful) to them. In this way, however, pragmatists can immediately recognize what is useful and what is not, in spite of their wrong perception of colors. Thus, they are evolutionarily favored over realists, whose perceptual systems offer a more accurate picture of things, but which need time and effort to collect a more detailed chromatic information and to compute from it whether a given thing is useful or not. A model in evolutionary game-theory set up by Hoffman and collaborators showed then that pragmatists would flourish and realists would be driven to extinction.

If this were true, evolution would favor useful but false perception, and this would mean that our own perceptual representations of the world are largely wrong (this is also argued by Stich (1991: Ch. III)). This, by the way, might suggest that we should largely discard the manifest image and rely on theoretical science for a more precise picture of physical reality. Moreover, evolutionary epistemology could no longer support both commonsense realism and scientific realism by arguing that true perceptual beliefs are favored by evolution, and philosophical skepticism would gain momentum.

In their paper, however, Angelucci and colleagues argue that the above study failed to consider environmental modifications: when conditions change, differences which were previously irrelevant to utility may become relevant. For instance, it may become the case that all and only green things are useful. In this way, pragmatists would become utterly confused, still "believing" that certain blue things (appearing grey to them) are useful and that certain green things (appearing brown to them) are not. Thus, they would soon become extinct, while realists would readily adapt to the new conditions, because they can properly distinguish green from other colors. To press their point, they propose a different model, incorporating the effects of environmental change, showing that in this model organisms able to produce more realistic representations of the world are favored in the long run.

Of course, these kinds of models are necessarily quite idealized, and their scope is limited by the particular assumptions incorporated. They are rather like particular thought-experimental settings. For instance, much depends on whether environmental evolution is discontinuous, with prolonged periods of stability between one change and the next, or it is ubiquitous and continuous: in the latter case, it seems, realists would always be ahead of pragmatists.

Therefore, while a model like the present one cannot warrant too general and certain conclusions, at least it suggests that evolutionary game theory might not bring so grim news for the perceptual accuracy of the manifest image, after all. In fact, it might be observed that the various species of the genus *homo* were distinguished from other animals precisely by their flexibility and ability to exploit even minor changes. Even more importantly, they didn't wait for the environment to modify the utility functions, but they always actively changed them by "inventing" ever new ways to take advantage of the environment. From this point of

view, it might be argued that perceptual realism has been one of the distinctive features of our species, and one of the keys to its evolutionary success.

References

Alai, M. 2014a, "Novel Predictions and the No Miracle Argument", *Erkenntnis*, 79, 2, 297-326.

Alai, M. 2014b, "Defending Deployment Realism against Alleged Counterexamples", in Bonino, G., Jesson, G., and J. Cumpa (eds.), *Defending Realism: Ontological and Epistemological Investigations*, Boston-Berlin-Munich: De Gruyter, 265-90.

Alai, M. 2019, "The Underdetermination of Theories and Scientific Realism", *Axiomathes-Epistemologia*, 29, 6, 621–37.

Bokulich, A. 2008, *Reexamining the Quantum–Classical Relation – Beyond Reductionism and Pluralism*, Cambridge: Cambridge University Press.

Buzzoni, M. 2008, *Thought Experiment in the Natural Sciences: A Transcendental-Operational Conception*, Würzburg: Königshausen & Neumann.

Carnap, R. 1928, *Der Logische Aufbau der Welt*, Leipzig: Meiner Verlag.

Cevolani, G., Crupi, V., and Festa, R. 2011, "Verisimilitude and Belief Change for Conjunctive Theories", *Erkenntnis*, 75, 183-202.

Cevolani, G., Festa, R., and Kuipers, T.A.F. 2013, "Verisimilitude and Belief Change for Nomic Conjunctive Theories", *Synthese*, 190, 3307-24.

Churchland, P.M. 1981, "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy*, 78, 2, 67-90.

Churchland, P.M. 1988, "Folk Psychology and the Explanation of Behaviour", *Proceedings of the Aristotelian Society*, Supp. Vol. 62, 209-21.

Eddington, A.S. 1928, *The Nature of the Physical World*, New York: The Macmillan Company.

Giere, R.N. 1988, *Explaining Science: A Cognitive Approach*, Chicago: Chicago University Press.

Hacking, I. 1983, *Representing and Intervening*, Cambridge: Cambridge University Press.

Hempel, C.G. and Oppenheim, P. 1948, "Studies in the Logic of Explanation", *Philosophy of Science*, 15, 2, 135-75; reprinted in Hempel 1965.

Hempel, C.G. 1965, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Hoffman, D.D., Singh, M., and Mark, J.T. 2013, "Does Evolution Favor True Perceptions?", in Rogowitz, B.E., Pappas, T.N., and de Ridder, H. (eds.), *Proceedings of the SPIE 8651, Human Vision and Electronic Imaging*, XVIII, 865104.

Hoffman, D.D., Manish, S., and Prakash, S. 2015, "The Interface Theory of Perception", *Psychonomic Bulletin & Review* 22, 6, 1480-1506.

Husserl, E. 1936, *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: Eine Einleitung in die phänomenologische Philosophie*, The Hague: Martinus Nijhoff, 1954. Engl. Transl.: *The Crisis of European Sciences and Transcendental Phenomenology*, Evanston: Northwestern University Press, 1970.

Knuuttila, T. 2021, "Epistemic artifacts and the Modal Dimension of Modelling", *European Journal for Philosophy of Science*, 11, 65.

Kuipers, T.A.F. 2000., *From Instrumentalism to Constructive Realism*, Dordrecht: Springer.

Laudan, L. 1981, "A Confutation of Convergent Realism", *Philosophy of Science*, 48, 19-49.

Main, J., Weibusch, G., Holle, A., and Welge, K.H. 1986, "New Quasi-Landau Structure of Highly Excited Atoms: The Hydrogen Atom", *Physical Review Letters*, 57, 2789-92.

Musgrave, A. 2006-2007, "The 'Miracle Argument' for Scientific Realism", *The Rutherford Journal*, *The New Zealand Journal for the History and Philosophy of Science and Technology*, 2, http://www.rutherfordjournal.org/article020108.html.

Poincaré, H., 1905, *La Science et l'Hypothèse*, Paris: Flammarion.

Polanyi, M. 1958, *Personal Knowledge: Towards a Post-Critical Philosophy*, London: Routledge and Kegan Paul; quotes from II ed. 1962.

Polanyi, M. 1969, *Knowing and Being*, London: Routledge & Kegan Paul.

Psillos, S. 1999, *Scientific Realism: How Science Tracks Truth*, New York: Routledge.

Schlick, M. 1938, *Form and Content: An Introduction to Philosophical Thinking*, in Schlick, M. *Gesammelte Aufsätze*, 1926-1936, Wien: Gerold, 151-249.

Sellars, W. 1962, "Philosophy and the Scientific Image of Man", in Colodny, R.G. (ed.), *Frontiers of Science and Philosophy*, Pittsburgh: University of Pittsburgh Press, 35–78.

Sellars, W. 1963, *Science, Perception and Reality*, London: Routledge & Kegan Paul and New York: The Humanities Press; reissued in 1991 by Ridgeview: Atascadero.

Stanford, K. 2006, *Exceeding Our Grasp: Science, History and the Problem of Unconceived Alternatives*, Oxford: Oxford University Press.

Stich, S. 1983, *From Folk Psychology to Cognitive Science*, Cambridge, MA: MIT Press.

Stich, S. 1991, *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*, Cambridge, MA: MIT Press.

Van Fraassen, B. 1980, *The Scientific Image*, Oxford: Oxford University Press.

Wittgenstein, L. 1921, "Logisch-Philosophische Abhandlung", in *Annalen der Naturphilosophie*, 14.

Zamora Bonilla, J. 1992, "Truthlikeness without Truth: A Methodological Approach", *Synthese*, 93, 343-72.

Zamora Bonilla, J. 1996, "Verisimilitude, Structuralism, and Scientific Progress", *Erkenntnis*, 44, 25-47.

Conceiving the Inconceivable: An Assessment of Stanford's New Induction

Giovanni Buonocore, Emilia Margoni,***

*Francesca Pero**

* *University of Florence*

** *University of Pisa, University of Florence, Université de Genève*

Abstract

Stanford's unconceived alternative argument is inductively based on the history of science and tells us that when a scientist is choosing a theory T_1 at time t_1 over a set of less promising alternatives, she is concurrently failing to conceive valid theoretical alternatives to T_1, i.e., theories that will be accepted by a scientific community at later times, thus displacing T_1. The aim of the present paper is to argue that the actual strength and reach of Stanford's argument sensibly vary according to the status of the unconceived alternatives at time t_1, i.e, whether they are *conceivable* (theories that could be conceived by scientists at t_1, but in fact are not) or *inconceivable* (theories which can not be conceived at t_1 as they are incompatible with scientists' background knowledge at t_1). As Stanford does not explicitly address this issue, we give reasons to conclude that alternatives considered in the unconceived alternative argument are supposedly conceivable at time t_1, and we investigate the consequences of this conclusion for the alleged novel induction the argument draws upon. We then investigate what are the implications for Stanford's analysis if inconceivability is considered as a possible status of an unconceived alternative at t_1, and we argue that, in this case, Stanford's antirealism has to be severely restricted to specific phases of theory-change, thus making room for tamed forms of realism.

Keywords: Instrumentalism, Pessimistic induction, Scientific realism, Unconceived alternative argument.

1. Introduction

Realist and antirealist stances have been developed into such articulated proposals that defining the key features of both positions risks ending in deadlock. Broadly speaking, scientific realism is taken as a "positive epistemic attitude" (Chakravartty 2017) towards the content of well-established scientific theories and models, whereas antirealism either questions the even approximate truthfulness of currently available scientific paradigms or declares to be agnostic about it.

Argumenta 9, 1 (2023): 27-41
ISSN 2465-2334

© 2023 G. Buonocore, E. Margoni, F. Pero
DOI 10.14275/2465-2334/202317.buo

Stanford's (2006) proposal sets forth novel arrows in the quiver of the antirealist party. To this end, he develops the unconceived alternatives (UA) argument that supposedly provides a new form of induction (NI), directed towards theorists rather than theories, as opposed to the old pessimistic induction (PI). The core of the UA argument is the claim that scientists have repeatedly failed to conceive of reasonable alternatives to time by time well-established scientific theories. On this view, the acceptance of a later theory provides the retrospective evidence of the inability to conceive of at least one alternative at the time the earlier theory was conceived, and this is what shall be inductively projected to current successful theories. Coupling this historical consideration with the assumption that scientific practice typically operates via eliminative inference methodology, it looks as if scientists are not able to exhaust the space of alternative theoretical explanations for a given set of phenomena.

According to Stanford, the problem of UA is different from the one posed by classical underdetermination from empirical equivalents insofar as it refers to the very epistemic, cognitive limits of those human agents—the scientists—who are in charge of delving into the maze of plausible theoretical candidates for a given set of phenomena (Stanford 2006: 16-17). Put it otherwise, Stanford's proposal to shift the focus from theories to theorists is what, in his intention, should strengthen the antirealist argument. It is very unlikely, he reasons, that current scientists have succeeded in what their predecessors have failed to, namely, to exhaust the space of plausible alternative theories for a given set of phenomena. On this view, for all available evidence at every given moment and in every sociocultural context there are always unconceived alternatives. And these unconceived alternatives are what, according to Stanford, seriously undermine scientific realism broadly construed.

Stanford's (2006) UA argument has been discussed and criticized by various authors. Based on diverse argumentative lines, they have pinpointed a few weaknesses that might affect his theoretical framework. On the one hand, Saatsi (2009, 2019) argues that Stanford's new induction does not look that novel at all, in that the problem of UA is still based on the well-known traditional weaponry of antirealists, namely, the underdetermination problem and the pessimistic induction. On the other hand, both Winther (2009) and Rowbottom (2019) claim that Stanford's proposal is too focused on theories and should rather extend the analysis to other aspects of scientific practice, to offer a more appropriate characterization of the latter.

In this paper we argue that Stanford's proposal does not take into consideration a crucial distinction, i.e., the one between conceivable and inconceivable alternatives. Conceivable alternatives are those theories which could have been conceived by scientists involved in a certain field at a specific time but in fact were not, despite these alternatives' compatibility with the evidence available at that time and the context's background assumptions. Conversely, inconceivable alternatives are those theories whose conceivability is prevented by empirical, methodological, and theoretical limitations.[1] Of course, being theories at stake (and

[1] It may be argued that the notion proposed here of inconceivable alternatives conflicts with the fact that there are historical cases in which a certain theory is developed despite being formally inconceivable. However, our argument is not to deny that scientists can conceive of alternative theories even in the absence of theoretical, methodological, and empirical elements—a paradigmatic example being the formulation of the heliocentric

assuming we are not extending platonism to theories themselves and treat them as abstract, non-mental, objects), we are considering conceivability and inconceivability as relative concepts, i.e, predicated in relation to an individual or an epistemic community who is capable (or incapable) of formulating the theory.

On the backdrop of this distinction, we argue that, whether an unconceived alternative at t_1 is conceivable or inconceivable, Stanford's argument and its anti-realist claim are affected. We reach this conclusion by first arguing that the status of an unconceived alternative at time t_1 that is consistent with Stanford's UA argument is conceivability, mainly for two reasons. First, were UA inconceivable at t_1 it would be hard to maintain the very core of Stanford's argument: how could we talk of a failure in conceiving something if it was, in fact, inconceivable? In other words, as we are going to argue in the following section, for Stanford's argument to even kick off and be a real threat to scientific realism (and a compelling form of the underdetermination argument against it) conceivability, as a feature of unconceived alternatives, should be as relevant as the other explicitly considered by Stanford.

Second, we think that the very requirements Stanford sets up for an unconceived alternative strongly hint at conceivability as its status at t_1. These requirements mainly amount to being empirically non-equivalent yet equally well confirmed by evidence as the 'dominant' theory and, more importantly, to be compatible with the same constraints and general metaphysical principles that guided the development and acceptance of the 'dominant' theory. According to our reading, Stanford presents unconceived alternatives as *consistent with* (which we read as *conceivable according to*) the evidence and the scientific environment (methodological and metaphysical) at stake. Therefore, according to Stanford's analysis and the conceivability of UA that, we argue, it calls upon, the only elements preventing the conception of UA are the scientists' epistemic limits. We agree with Stanford that scientists, as epistemic agents, come equipped with a remarkable yet limited capacity to explore the complex targets of scientific research. However, such limitations are mainly imposed by the background knowledge collected up to the time scientists live in. Consequently, in such a conceptually bounded context, many of the alternatives later accepted by a different or following scientific community could not even be imagined and formulated, let alone conceived. In other words, Stanford portrays as a limit, namely as an epistemic failure, what is not even attainable in principle, i.e., conceiving something that we will probably not be even able to adopt for explanatory purposes, as it clashes with the way we conceptualize physical reality. We attempt to make this point by considering a case study where a formerly unconceived alternative was first and foremost inconceivable (Sect. 3).

We then turn to investigate what kind of notion of conceivability would be consistent with Stanford's account and why such a notion would undermine the core of the NI, being it only allegedly a property relative to theorists' capabilities, and actually pertaining to theories. If this is the case, the focus of the meta-inductive claim would be on theories rather than on theorists, just as in the old PI (Sect. 4). We finally investigate what are the implications for Stanford's accounts and

theory by Aristarchus of Samos despite Aristotelian mechanics and the lack of empirical data. What we mean to question is that one can construct a novel antirealist argument based on Stanford's shift from theories to theorists. And this, we believe, gets signaled by focusing on the distinction between conceivable and inconceivable alternatives.

the reach of its antirealist stance if inconceivability is considered as a possible status of an unconceived alternative at t_1 and we argue that, in this case, Stanford's antirealism has to be restricted to those specific phases of theory-change in which a certain theory remains unconceived, despite being conceivable (Sect. 5).

2. Conceivable *vs* Inconceivable Theories

In considering the spectrum of theoretical alternatives to a given set of phenomena, we would like to introduce a taxonomy that will pave the way for the present discussion. Our claim is that the distinction between conceivable and inconceivable alternatives is not only valuable, but also necessary if one aims to take at face value the import of Stanford's proposal. Our theoretical framework envisages the following taxonomy:

(1) inconceivable – unconceived
(2) conceivable – unconceived
(3) conceivable – conceived

This paper will focus on the problematic relationship between options (1) and (2). Our claim is that the difference between (1) and (2) is crucial when it comes to the problem of UA. Missing this distinction, our argument runs, risks rendering Stanford's argument either trivial or incomplete. For it is one thing not conceiving alternative theories that we could have conceived—being them conceivable—but did not for some (putative) socio-cultural limitations, quite another not conceiving those alternative theories which are inconceivable because of empirical, methodological, and theoretical limitations.

Stanford's historical reconstruction draws a picture in which theories which are newly developed repeatedly turned out to be the unrecognized, unconceived alternative to an antecedent well established one, with the intent to inductively generalize such pattern to possibly every case of theory-change. This inductive generalization is what determines his antirealist stance towards scientific practice broadly construed, rather than confining his analysis to specifically selected theories or specifically selected theory-change contexts.

However, Stanford's analysis of historical records makes no reference to the concept of inconceivability and seems to only account for alternatives of type (2)—namely, conceivable theories that remained unconceived because of a contingent failure within the scientific community to conceive them. According to Stanford, such a failure is due solely to the fact that "our cognitive constitutions or faculties are not well suited to exhausting the kinds of spaces of serious candidate theoretical explanations" (45). In fact, Stanford devotes part of his book to the analysis of cases from the history of science where there is an alleged continuity among evidential, metaphysical, and methodological constraints, such as in the case of the transition first from Darwin's pangenesis theory of inheritance to Weismann's germ-plasm theory, then to the Mendelian theory and, finally, to contemporary molecular genetics.[2] Such a continuity would rule out any potential incompatibility of background knowledge and assumptions among scientific communities as a possible reason for scientists' failure to conceiving "scientifically serious alternatives" to a

[2] It is important to note that what Stanford (2006) means by "metaphysical limitation" remains unspecified in the context of theory generation. In the following section we thus adopt the more adjustable notion of "theoretical limitation".

theory T_1. Such a label is particularly relevant to Stanford's analysis as it marks the difference between the unconceived alternatives and those alternatives invoked by the traditional form of underdetermination. In fact, the alternatives considered by the UA argument are not "construct[ed] parasitically so as to perfectly mimic the predictive and explanatory achievements of our own theories" (18-19), but are genuine theoretical alternatives that simply remain unconceived up to a certain time and eventually "accepted by some actual scientific community" (21). More importantly, these alternatives were "scientifically serious even by the standards of the day despite being unconceived and therefore unconsidered by theorists at the time" (60). This is exactly what should make the UA argument a bigger threat for scientific realism than the traditional underdetermination argument: being the alternatives it considers scientifically serious, UA cannot be reduced and dismissed by a realist as a philosophical speculation valid only from the logical viewpoint, or in-principle.

Summing up, the main claims of Stanford's UA argument are the following:

(a) At a time t_1, a theory T_1 is conceived and preferred over a set of other conceived but not equally well-confirmed theories.
(b) At a later time t_2, an empirically non-equivalent, but equally well confirmed, alternative T_2 is conceived and preferred over T_1.
(c) At the time t_1, the theory T_2 was conceivable (as equally well confirmed) despite remaining unconceived until t_2.

Now, let us make the conceivability condition explicit:

(L1) T_2 needs to be at least equally well supported by the evidence that supported T_1, at t_1, and compatible with T_1's background assumptions (or constraints), whether they be theoretical, empirical, or methodological.[3]

The question that naturally arises is: does the UA argument require new theories to be conceivable at the time in which old ones were conceived and accepted? If this were the case the choice between T_1 and T_2 would be underdetermined at t_1, precisely when the conceivability condition of T_2 needs to be met. As we shall see in the following section, the implications of such a framework are quite radical and implausible when generalized and applied to other (well-known) cases of theory-change, for which the conceivability condition clearly does not hold. In fact, what happens between t_1 and t_2 matters. As pointed out in Magnus' critical assessment of the UA argument (2006), a theory T_2 is conceived within a period of revolution to try to account for some evidential anomalies the theory T_1 struggles with. During the period of controversy, T_1-supporters try to account for such anomalies within T_1 itself, while others formulate the new theory T_2. At this moment (and only at this moment) there really is a problem of underdetermination between T_1 and T_2, but as further, decisive, evidence is gathered, the problem might eventually vanish, thus defining a preference between T_1 and T_2.

Let us now turn our attention to a paradigmatic case in which this aspect is brought up to the forefront.

[3] The issue of background assumptions and their role in theory choice is a topic which exceeds by far the limited scope of the present paper. We just like to note in passing that the ambiguities associated to this topic have been tackled by eminent scientists (Einstein 1936), philosophers of science (Reichenbach 1958, Kuhn 1970) and have a proper status as an issue in social epistemology (Longino 2002, Nelson 1993, Potter 1996).

3. From Newtonian Mechanics to Relativistic Physics: A Case Study

The UA framework, while successfully applicable to the instances of scientific theorizing from biological sciences Stanford considers, is not as successful when, e.g, applied to the history of physics. This is rather evident when considering the paradigmatic passage from Newtonian theory of space and time to Einstein's theory of special relativity (STR). As Stanford states, "the evidence available at the time the earlier theory was accepted offered equally strong support to the (then-unimagined) later alternative" (19). As shown at the end of the previous section, the UA argument implicitly requires the conceivability condition of the later theory to be met already at the time the earlier was conceived and accepted: at the time of Newton there were no empirical, theoretical, or methodological constraints that could prevent the scientific community to conceive of STR, but only cognitive-epistemic limitations. STR was conceivable and, yet, remained unconceived until 1905.

Drawing on the works by DiSalle (1990), Norton (2004) and Cassini and Levinas (2019), in the following we trace back those theoretical, empirical, and methodological constraints for STR's conceivability that, contrary to Stanford's point, were in fact inaccessible at the time of Newton and that turned out to be essential to Einstein's fundamental intuition about the relativization of the notion of simultaneity.

As for the theoretical constraints, conceiving an equivalence-class as the fundamental spatiotemporal framework required a level of abstraction attainable only with the mathematics of the 19th century. Thomson's (1884) reassessment of the laws of inertia highlighted the fundamental relation between Newton's laws of motion and inertial frames, namely, the existence of (at least) one inertial frame, with respect to which any other is in uniform motion. The point was that any inertial frame could be constructed as the "absolute" space in which all the others are uniformly moving, and, therefore, the crucial issue was no longer to identify the frame of reference in which the dynamical laws hold, but, rather, how the laws of motion are able to define an appropriate class of reference frames (DiSalle 2020: 23). Lange (1885), independently of Thomson's work, introduced a new definition of inertial system based on the intuition that all motion is relative: an inertial system is a coordinate system with respect to which three free particles move in straight lines and travel mutually proportional distances as they are projected from a single point and are moving in non-coplanar directions (DiSalle 1990). According to the laws of inertia, any fourth free particle will move uniformly with respect to any inertial system; thus, Newton's notion of absolute acceleration (and rotation) can be replaced by that of acceleration (and rotation), relative to an inertial system (and timescale). Although Lange's and Thomson's direct influence on Einstein, as well as their broader historical impact, is difficult to assess (DiSalle 1990: 140), by the beginning of 1900 the notion of inertial system had permeated the debate around mechanical philosophy and was assumed as the foundation for classical mechanics. In fact, Einstein (1905) took it for granted that his readers consider an equivalence-class of frames of reference rather than a privileged frame (see DiSalle 2020).

Turning to the empirical constraints necessary for STR to be conceived, the historical record of Einstein's oral presentations shows some explicit references to the relevance of Fizeau's results (see Shankland 1963: 48), although not stated

in published or unpublished works (see Norton 2004).[4] Fizeau tried to measure the relative speed of light in water, using a particular interference system that measured the effect of the moving medium on the speed of light itself, by observing interference fringes produced by two rays of light passing through two parallel pipes filled with water flowing in opposite directions.[5] Fizeau considered three hypotheses, only one of which to be confirmed by his experiment: (1) the ether has no interaction with the moving medium, (2) it is partially dragged by the moving medium (Fresnel's hypothesis), (3) it is fully dragged. He erroneously considered his observations of small fringes displacement to confirm (2), by assuming a portion of the ether was fixed to the water molecules, but Fizeau never considered that the effect could have been explained without any reference to matter-ether interaction (Patton 2011: 215). And, in fact, Lorenz (1895) considered this fourth hypothesis, and proved it to be the right one: the effects obtained by Fizeau, despite being compatible with (2), were determined solely by the reflection and refraction of light waves, rather than matter-ether interaction. This fact alone, however, did not prompt the Dutch scientist to abandon 'still ether' as a reference frame. It was only with the successive reinterpretation of Fizeau's experiment under the new conceptual framework of the equality of all inertial systems that its results turned out to be crucial for STR's conceivability.

Finally, as also pointed out by Norton (2004), Einstein's methodological debts to the writings of Hume and Mach are evident when it comes to his account of the nature of concepts in general rather than the specific analysis of space and time carried out by the two authors. Einstein himself pointed out that his intuition came from a reconsideration of certain types of concepts that physical theories include, which, in order for them to represent something physical, must be grounded in experience (Einstein [1917] 1954: § 8). Einstein (1916) makes explicit reference to the valuable method of conceiving concepts as physically meaningful only in so far as they are empirically grounded. But in Mach's writings specifically (see, e.g., Mach [1907] 1960), it also emerges a radical attitude towards fictional concepts that leads to their complete elimination from any relevant account of the physical world to which Einstein was reluctant (Holton 1968: 231). Hume's analysis ([1748] 1988) is also based on certain notions ("ideas") that must be grounded in sense experience ("impressions"), in line with Mach's empiricism. But on the other hand, Hume did not propose to completely eradicate such notions that were not empirically grounded, as in the case of causality. And, indeed, the reconceptualization of a fictional concept whose uncertain character is recognized but accommodated within the physical theory in such a way to "preclude unwitting introduction of false presumptions" (Norton 2004: 3) is precisely the theoretical step that Einstein took towards the relativization of the notion of simultaneity. It is perhaps for this reason that Einstein firsthand declared Hume's work having "much more influence" than Mach in the formulation of STR (Einstein 1949, as quoted in Norton 2004: 2).

4. What if Unconceived Alternatives Are Conceivable

The old induction statement is confined to theories and, in particular, it casts doubts on the truth of theoretical claims. Differently, NI redirects the pessimism

[4] For additional references on the influence of Fizeau's results see Einstein 1923 and Moszkowski 1972.

[5] For a detailed presentation of the experiment see Patton 2011 and Cassini and Levinas 2019.

of PI from theories to theorists as cognitive agents, asserting the impossibility for theorists to ever exhaust the space of possible alternatives to the theory accepted at a certain moment. According to Stanford, this sort of pessimism is difficult not to subscribe to, thus making NI a bigger draw for an antirealist than the old PI. In fact, Stanford claims, we have collected throughout history of science enough evidence to inductively rule out the possibility that future scientific communities will epistemically improve to the point they will not fail to exhaust the alternatives' space.

Stanford's analysis is convincing as long as the only possible scenario that leads to what he defines as a "scientifically serious alternative" (2006: 20), i.e., a theory later accepted by a scientific community, is the one depicted in condition (2) (Sect. 2): alternatives remained unconceived are formerly conceivable ones. We question whether we can talk of an epistemic failure in not conceiving a theory which then turned out to be a serious scientific alternative if the latter was inconceivable at the time another one was accepted. A way to approach such an issue is by conditionally investigating why Stanford would neglect the crucial difference between conditions (1) and (2) above. In fact, such an omission seems to contravene the interest Stanford proclaims for the "empirical exploration of the various dynamical processes that help explain how and why particular unconceived alternatives remain unconceived by particular (human!) scientists and scientific communities" (2009: 381).

One reason why Stanford could reject the distinction is built into the transient nature of the underdetermination as intended by NI, upon which the conceivability notion depends.[6] Recall that recurrent transient underdetermination requires unconceived alternatives to be equally (roughly, at least) well confirmed by the available evidence although empirically non-equivalent to their rival, and to be so up to a certain historical development when enough evidence has been collected so that the rival theory is differently confirmed. Such a requirement, Stanford claims, "deflects any suggestion that such alternatives were ignored on evidential grounds rather than simply unconceived" (2006: 26). Deflecting this sort of suggestion is crucial since it is a threat to the notion of conceivability: if the alternative unconceived theory is unable to make evidence intelligible, then the alternative is inconceivable.

The scientific context, together with its methodological, theoretical, and metaphysical assumptions, informs the way a scientific community of a certain time classifies phenomena and, consequently, collects evidence to test a hypothesis. This state of affairs negatively affects the idea of conceivability Stanford promotes as an atemporal quality of a theory.[7] In fact, inconceivability could be denied as a former status of an unconceived alternative by claiming that the very existence of evidence equally supporting the unconceived alternative suffices to make it

[6] Another reason why one could neglect the case unconceived alternatives were formerly inconceivable is because such occurrence is considered as impossible. We assume this would be too bold of a claim to subscribe.

[7] Magnus (2010) takes "being a scientifically serious alternative" to be treated by Stanford as a "timeless property of a theory" (7). We think that ascribing timelessness to the property of "being conceivable" is more consistent with Stanford's proposal. This is because "being a scientifically serious alternative" depends upon a subsequent and ultimate act of a scientific community which takes place at a specific moment in time while conceivability is something that, according to Stanford, could be predicated of a serious scientific alternative prior to its acceptance by a scientific community.

conceivable, independently of the theorist's epistemic ability to use that theory to read such evidence. In other words, as long as a theory is equally well supported by evidence as its rival(s), such theory has the ability to be conceived, regardless of whether the contemporary scientific community has the ability to conceive it. Consequently, conceivability has little to do with scientists' epistemic possibilities in that it is rather a way to sort out the possible interaction of theories alone with evidence as it is given by itself, and not accordingly to a scientific frame of reference. Evidence could be read in the light of the unconceived alternative, thus making it conceivable. However, the consistency of the later-accepted theory with respect to the evidence available at the time the later-overturned theory was dominant can be identified only retrospectively. Beforehand, it might be the case that, to the eyes of the scientific community of the time, any reading of the evidence according to standards incompatible with their epistemic, metaphysical or methodological background assumption was, in fact, impossible.

Neglecting that unconceived alternatives can be inconceivable is legitimate only if we narrow the analysis, and the inductive generalization we want to make with it, to theories as final byproducts of theorizing, as well as to their relations to evidence and to preceding accepted theories. The fact that scientists and scientific communities consistently fall short of conceiving scientifically serious alternatives at a certain time does not affect the conceivability of those theories at that time (being them conceivable regardless of humans' ability to conceive them) In fact, conceivability is cast in Stanford's analysis mainly as a property of theories, due to their relation to evidence rather than to a theorist's epistemic capacity. Accordingly, the predicament NI is about does not concern the disadvantaged epistemic position we are doomed to occupy across the history of science, that is, the position where serious scientific alternatives remain unconceived by us. Rather, the predicament is just the same that serves as the empirical premise for the old PI: the recurrent turnover of older theories in favour of new ones. Adding conceivability to the picture grants no novelty to the inductive argument, as the way Stanford casts this notion is not informative about the processes that lead unconceived alternatives to remain unconceived by certain scientific communities. Evidence could be read in the light of the unconceived alternative, thus making it conceivable.

So far, the novelty of NI has been questioned by looking at its inductive basis rather than at the concepts it is built upon. In fact, Magnus (2006) and Saatsi (2009) worry that NI either fails as induction or it is old-fashioned after all. They question Stanford's account of what might happen at time t_1 among rival theories (including unconceived ones), theorists and available evidence on the grounds of counterfactual claims, based on the concept of plausibility. Given that a scientific community comes with some standards of plausibility, which in turn fix some criteria for the definition of what counts as experience, it can be the case that a scientifically serious alternative, were it conceived and presented by any member of the community, would not have seemed to be plausible to the rest of it. Considering the stand-off between classical mechanics and relativity, the question arose whether the latter would have been considered as plausible had it been presented to the scientific community subscribing the former. More precisely, could the data available and, most of all, the way they were collected and interpreted by Newtonians, license any plausibility claim about the alternative reading provided by the theory of special relativity? Given that standards of plausibility change according to the scientific context

at stake and crucially determine what is classifiable as an experience, Saatsi and Magnus opt for a negative answer to that question.

Stanford (2009, 2017) replies to Magnus and Saatsi that the change of standards of plausibility across history of science does not suffice to conclude that genuine alternatives never existed and ever won't and that, consequently, NI is undermined. For this to be the case, it should also be assumed that implausibility actually prevented Newtonians from conceiving special relativity and also that future scientific communities will not undergo the same changes of scientific background assumptions, thus eventually discrediting previously entrenched judgments of implausibility. Therefore, *mutatis mutandis*, according to Stanford (2009) implausibility shows that we cannot rely on our own standards of plausibility and, consequently, we are doomed to occupy the same epistemic predicament of earlier scientific communities.

There is a crucial conceptual difference between the concepts of implausibility and inconceivability. Implausibility assumes as very unlikely the case where a theory which later turned out to be a serious scientific alternative was first conceived by someone in the scientific community and then immediately withdrawn as implausible. However, implausibility does not rule out such a case as impossible. On the other hand, inconceivability does not allow for such a circumstance to take place: the ephemeral conception of an alternative theory that has to wait many more years to be accepted by a scientific community is not an option, as long as what is required to conceive it is incompatible with the background assumptions held at that time. Ruling out the possibility for such a scenario, inconceivability does not lay itself open to Stanford's reply that time-dependence of plausibility judgements proves that science is at any time unreliable and that dismissed possibilities were actually preferable: they were not conceivable in the first place!

5. Accepting the Distinction between Conceivable and Inconceivable Alternatives: Possible Consequences for the Realist *vs* Antirealist Debate

Newton and his contemporaries did not simply fail to conceive of an alternative theory to Newtonian mechanics such as the theory of special relativity: in that context, the latter was in fact inconceivable. That being so, what is at stake is whether the inconceivability of a certain theory provides elements in favour of the antirealist perspective.

Now, to evaluate this point, let us unpack what the inconceivability of a theory stands for. As already discussed in the previous sections, if a theory is inconceivable because of empirical, methodological, and theoretical reasons, this means that, at the time the inconceivability is met, there is a lack of empirical data, or the absence of a suitable methodological apparatus, or the unavailability of an appropriate theoretical formalism.[8] Either way, two considerations can be made.

First, it is not simply the case that the community of scientists fail to conceive of an alternative theory for a given set of phenomena. Rather, the missing background knowledge is what prevents the attempts in the first place. Again, it looks

[8] Let us for now grant the (highly implausible) thesis according to which there simply are empirical data. We will come back to this point later on in this section.

as if the most we can concede to Stanford is that we are back to the old pessimistic induction situation. In other words, if a theory is unconceived because it is inconceivable, then Stanford's NI does not look that novel at all.

Second, there are different versions of realism and antirealism: a contemporary realist would hardly claim that we should accord currently successful theories a state of complete truthfulness. For in its broadest characterization, realism is simply taken as a positive attitude toward the content of well-established scientific theories and models (Chakravartty 2017). Obviously, then, those scientific theories and models may not accommodate lacking data, methodologies, and formalisms. If this is the case, then the problem is not that in each historical context scientists fail to exhaust the space of plausible alternatives to a given set of phenomena. Rather, the fact is that the set of phenomena (plus the associated methodological and theoretical toolkit) is insufficient to make a certain theory conceivable. But then a perhaps mild realist does have some elements to resist the antirealist claim of Stanford's UA argument. Indeed, she might claim that, though we currently lack those empirical, methodological, and theoretical elements that are necessary to the formulation of a currently inconceivable theory, there are good reasons to believe—*contra* Stanford's historical record—that when the conceivability condition is met, the theory gets eventually formulated. To conclude, if we confine our analysis to the first case, namely to a theory which is unconceived and inconceivable, it looks as if we either get back to the old pessimistic induction situation or that the realist party can cope with the inconceivability condition by smoothing her own commitments, thus conveying a form of "tamed" realism.

Let us now turn our attention to the subtler case in which a certain theory is conceivable, yet unconceived. To stick to the example of Sect. 3, this condition applies to that period, between the end of the 19th century and the beginning of the 20th century, in which both the Fizeau and the Michelson-Morley experiments were being discussed and Lorentz provided a mathematical formalism and a theoretical hypothesis for that experiment. Again, the point at stake is to evaluate the consequences of a theory being conceivable, yet unconceived upon the realist *vs* antirealist debate. This case looks more hospitable to Stanford's proposal, in that one cannot advocate the inconceivability condition whereby the mild realist is able to construct her counterargument.

We acknowledge that Stanford's proposal does apply to those historical cases in which a certain theory is conceivable-yet unconceived, with the following provisos. First, if one confines Stanford's proposal to such cases, it looks as if his overall enterprise has to seriously narrow its scope. Indeed, the main claim of the present paper is precisely to argue that Stanford's lesson cannot be generalized to every theory-change. Second, provided that Stanford's argument applies to conceivable, yet unconceived alternative theories, one of the most challenging points becomes how to identify them (on a similar vein, see Ruhmkorff 2019: 3937-38).[9] Finally, there is a third problem which, according to us, affects Stanford's proposal and that, in a way, might be used against us, for it implies problematizing even the conceivability *vs* inconceivability condition.

[9] Importantly, the selection of theories exposed to the problem of UA—namely, those who are unconceived yet conceivable—can only be reconstructed retrospectively. This is why, even if the scope of Stanford's proposal gets narrowed, it is still unclear to which case-studies it should actually be applied.

The point at stake is: How can we say that a certain theory is conceivable, yet unconceived? Evidently, to do so, we are hypothesizing that a certain set of empirical, methodological, and theoretical components can be mapped in a somewhat traceable way from one theory to another. It is no coincidence, then, that Stanford (2006: 22) declares his argument to be incompatible with the Kuhnian notion of incommensurability. However, there is no clear justification for such a statement. Better, as already noted by Winther (2009), on Stanford's account a whole perspective is lacking, to such an extent that data are given in an utterly unproblematic way. On this reading, if one aims at engaging with the realist *vs* antirealist debate, one should primarily investigate the bottom-up mechanisms whereby entities get reified in the scientific practice.[10]

The example of the transition from Newtonian to relativistic physics hardly serves as a case study supporting Stanford's view. And, more importantly, one needs to be very cautious in regarding data as preconceptual elements that can be selectively rearranged in various theoretical contexts. In fact, data must be interpreted as the by-product of scientific pragmatic agency, so much so that it does not make sense to evaluate them without reference to all the background assumptions, which in turn reify, i.e., produce those same data. In conclusion, the available evidence Stanford invokes to construct his argument does not look amenable to a straightforward mapping between different theoretical backgrounds. For if one is interested in detailing the scientific practice, one has to seriously engage with that myriad of factors (such as experiments, models, techniques, observations, tools, expectations, predictions) that not only figure in a certain scientific context but help shape it in the first place.

6. Final Remarks

Stanford's analysis is surprisingly elusive about the concept of inconceivability, despite its pivotal role for the UA argument. In particular**,** t**h**e issue of whether a certain theory is compatible with a scientific community's background assumptions—an issue which is required by the conceivability condition—is mentioned yet left untackled by Stanford. And this, we think, is a gap in his proposal that needs to be filled in, as making the conceivability condition explicit might improve the UA argument resilience to criticisms such as the one advanced here.

Stanford does acknowledge the possibility for a criticism in line with the one we focused on here, when he mentions the following question: "were the later alternatives unconceived by earlier practitioners really even serious ones at the time, given profound differences in available evidence, metaphysical presuppositions about nature, and methodological assumptions about its investigation?" (Stanford 2006: 59). However, we find his reaction to this specific point far from being a plausible answer, in that it does not really address the issue at stake. In fact, Stanford confines his answer to a mere hypothesis about the assumption that anyone raising the question above may hold, that is, that we currently occupy a privileged position in the history of science whose methodological assumptions and metaphysical presuppositions will not undergo the same fate of previously

[10] As a particularly instructive discussion of this problem, see Patton's (2011) analysis of Fizeau's classic optical experiments. According to her, there is no unequivocal way in which to account for the results of a certain experiment and thus there is no way in which the latter can be taken as unequivocally given.

discarded ones. That one needs to hold such an assumption to raise the problem of inconceivable unconceived alternatives is false advertising. Trivially, we can peacefully claim that the theory of special relativity could not be a serious scientific alternative for Newtonians, given the background assumptions they subscribed to, while maintaining that the theory of special relativity and everything it implies, from both a metaphysical and a methodological standpoint, will not be our final view on space, time, and matter.

We argued that it is not always the case that the failure to conceive theoretical alternatives happens across slices of history of science where there is a continuity in metaphysical, methodological, and empirical assumptions like the one Stanford envisages between mid-to-late 19th century theorists investigating inheritance and generation and contemporary theorists dealing with genetics and embryology. Even the most committed realists such as Saatsi (2019) concur that typically this continuity does not hold among theory-change. Provided that, also in the historical case Stanford considers, the employed notion of background assumptions remains unduly vague, we took the liberty to distinguish the sense in which Stanford uses it, i.e., as limitations that do not prevent conceiving, from the sense usually ascribed to the concept of background assumptions, unconceived alternative theories. Hence, we introduced the alternative label of "sociocultural limitations", for we think the latter is more consistent with the way Stanford employs the notion of background assumptions—whereas on the standard interpretation the latter would hardly accommodate cases in which radically different and subsequent theories were conceivable, yet the scientific community of that time failed to conceive them.[11]

A further remark concerns the sharp distinction Stanford seems to imply between theories and theorists, which patently clashes with the theory-ladenness of data collected to test theories as well as the empirical results that come out of this process. In section 4, we have given reasons to conclude that Stanford treats conceivability as an atemporal property of theories, rather than as a theorist's epistemic possibility with respect to a theory's content. As long as a theory is equally well supported by evidence as its rival(s), such theory has the ability to be conceived, regardless of whether the contemporary scientific community has the ability to conceive it. Put it otherwise, to assess whether a theory is conceivable we only need to look at its relationship to experience. The implication that conceivability is determined by the theory-evidence relation, with no inclusion of theorists as producers and consumers of theories, is hard to subscribe in general. In particular, it clashes with Stanford's project to redirect the pessimism inductively justified—and the antirealism thereof—to theorists rather than theories.

In the last section, we explored the consequences of accepting the distinction we highlight between conceivable *vs* inconceivable alternatives. As Stanford's main goal was to provide novel theoretical support to broadly anti-realist claims, we emphasized that, whenever a theory is unconceived (also) because of its inconceivability, then there are ways in which a tamed form of realism can be advocated. Still, we believe, the most interesting cases—at least when it comes to evaluate the UA argument—are those in which a certain theory is unconceived despite the absence of theoretical, empirical, and methodological impeding factors. For it is precisely in these cases that the untenability of the distinction

[11] Stanford also adopts the term "conceptual barriers or limitations" (132) to signal what might prevent from conceiving serious scientific alternatives to the accepted ones.

between theories and theorists comes to the forefront, while scaling back the realist *vs* anti-realist debate. In a way, we concur with Stein (1989: 56) when he argues that, though trying to unravel the role of theories in the evolving process of discovery is both intriguing and relevant, the matter in question cannot be simply resolved in terms of the realist *vs* instrumentalist discourse. Rather, we claim, one should focus on the way in which observations, models, predictions, methods, instruments, experiments, and values (Rowbottom 2019) mutually interact in the context of scientific practice and how other factors, such as ontological assumptions, theoretical principles, and standards of evidence play a crucial role in both theory-building and evidence assessments. Our claim is that to properly engage with the way in which scientific practice is carried out, one should primarily target those background assumptions that allow for the individuation of data and their selective arrangement within a specific theoretical context. This is what, in our view, is particularly wanting in Stanford's account and what prompted us to unravel the distinction between conceivable *vs* inconceivable alternatives in the first place.[12]

References

Chakravartty, A. 2017, "Scientific Realism", *The Stanford Encyclopedia of Philosophy*, Zalta, E.N. (ed.), https://plato.stanford.edu/archives/sum2017/entries/scientific-realism

Cassini, A. and Levinas, M.L. 2019, "Einstein's Reinterpretation of the Fizeau Experiment: How It Turned out to Be Crucial for Special Relativity", *Studies in History and Philosophy of Science*, *Part B: Studies in History and Philosophy of Modern Physics*, 65, 55-72.

DiSalle, R. 1990, "Conventionalism and the Origins of the Inertial Frame Concept", *PSA 1990: Proceedings of the 1990 Biennial Meeting of the Philosophy of Science Association*, 139-47.

DiSalle, R. 2020, "Space and Time: Inertial Frames", *The Stanford Encyclopedia of Philosophy*, Zalta, E.N. (ed.), https://plato.stanford.edu/archives/win2020/entries/spacetime-iframes

Einstein, A. 1916, "Ernst Mach", *Physikalische Zeitschrift*, 6, 29, 101-104.

Einstein, A. [1917] 1954, "Über die spezielle and die allgemeine Relativitätstheorie (Gemeinverständlich)", Braunschweig: Friedr. Vieweg and Sohn; 15th expanded edition, *Relativity: The Special and the General Theory*, London: Methuen.

Einstein, A. 1923, "How I Created the Theory of Relativity", *Jun Ishiwara's Notes on Einstein's Lecture at Kyoto University*, *Collected Papers*, 13, 399, 629-36; trans. 636-41.

Einstein, A. 1949, "Autobiographical Notes", in Schilpp, P.A. (ed.), *Albert Einstein: Philosopher-Scientist*, New York: Open Court, 1-94.

Holton, G.J. 1968, "Einstein, Mach, and the Search for Reality", in *Thematic Origins of Scientific Thought: Kepler to Einstein*, Cambridge, MA: Harvard University Press, 1973, 219-59.

[12] Francesca Pero and Emilia Margoni acknowledge financial support from MIUR through the PRIN 2017 project "The Manifest Image and the Scientific Image" (Prot. 2017-ZNWW7F-004).

Hume, D. [1748] 1988, *An Enquiry Concerning Human Understanding*, Chicago and La Salle: Open Court.

Kuhn, T.S. 1970, *The Structure of Scientific Revolutions* (3rd edition), Chicago: University of Chicago Press.

Lange, C.G. 1885, "Über das Beharrungsgesetz", *Berichte der Königlichen Sächsischen Gesellschaft der Wissenschaften zu Leipzig*, *Mathematisch-physische Classe*, 37, 333-51.

Longino, H. 2002, *The Fate of Knowledge*, Princeton: Princeton University Press.

Mach, E. [1907] 1960, *The Science of Mechanics: A Critical and Historical Account of Its Development*, 6th ed., trans T.J. McCormach from 9th Germ. ed., La Salle: Open Court.

Magnus, P.D. 2006, "What's New about the New Induction?", *Synthese*, 148, 2, 295-301.

Magnus, P.D. 2010, "Inductions, Red Herrings, and the Best Explanation for the Mixed Record of Science", *The British Journal for the Philosophy of Science*, 6, 4, 803-19.

Moszkowski, A. 1972, *Conversations with Einstein*, New York: Horizon Press.

Nelson, H.L. 1993, "Epistemological Communities", in Alcoff, L. and Potter, E. (eds.), *Feminist Epistemologies*, New York: Routledge, 121-59.

Norton, J.D. 2004, "How Hume and Mach Helped Einstein Find Special Relativity", in Friedman, M., Domski, M., and Dickson, M. (eds.), *Discourse on a New Method: Reinvigorating the Marriage of History and Philosophy of Science*, 359-86.

Patton, L. 2011, "Reconsidering Experiments", *The Journal of the International Society for the History of Philosophy of Science* 1, 2, 209-26.

Potter, E. 1996, "Underdetermination Undeterred", in Nelson, L.H. and Nelson, J. (eds.), *Feminism, Science, and Philosophy of Science*, Dordrecht: Kluwer, 121-38.

Reichenbach, H. 1958, *The Philosophy of Space and Time*, New York: Dover Publications.

Rowbottom, D.P. 2019, "Extending the Argument from Unconceived Alternatives: Observations, Models, Predictions, Explanations, Methods, Instruments, Experiments and Values", *Synthese*, 196, 3947-59.

Ruhmkorff, S. 2019, "Unconceived Alternatives and the Cathedral Problem", *Synthese*, 196, 3933-45.

Saatsi, J. 2009, "Review Symposium: Grasping at Realist Straws", *Metascience*, 18, 355-62.

Saatsi, J. 2019, "Historical Inductions, Old and New", *Synthese*, 196, 3979-93.

Shankland , R.S. 1963, "Conversations with Albert Einstein", *American Journal of Physics*, 31, 47-57.

Stanford, K. 2006, *Exceeding Our Grasp: Science, History and the Problem of Unconceived Alternatives*, Oxford: Oxford University Press.

Stanford, K. 2017, "Unconceived Alternatives and the Strategy of Historical Ostension", in Saatsi, J. (ed.), *The Routledge Handbook of Scientific Realism*, London: Routledge, Ch. 17.

Stein, H. 1989, "Yes, but. . . Some Skeptical Remarks on Realism and Anti-Realism", *Dialectica*, 43, 1-2, 47-65.

Thomson, J. 1884, "On the Law of Inertia; the Principle of Chronometry; and the Principle of Absolute Clinural Rest, and of Absolute Rotation", *Proceedings of the Royal Society of Edinburgh*,12, 568.

Winther, G. 2009, "Review Symposium: Grasping at Realist Straws", *Metascience*, 18, 370-79.

Structure Representation of Deep-Learning Models: The Case of AlphaFold

Giovanni Galli

University of Urbino Carlo Bo

Abstract

The scientific enterprise enriches the debate about models. In particular, in the field of structural biology, a new deep-learning neural network system called AlphaFold has been applied for many purposes. It allows us to predict a protein's structure with high accuracy. I will present the system in light of the discussion of structure representation and argue for a specific kind of representational relation holding between the predicted model structure and its target-system. By doing so, I will criticize the artifactual approach advanced by Knuuttila (2021) and present the features that characterize the predicted structures of AlphaFold as simulation models.

Keywords: Scientific representation, Deep-learning models, AlphaFold, Protein structure determination.

1. Introduction

The notion of model is one with a wide polysemy within the sciences and philosophy. There is no unique conceptual framework and definition able to define all the models involved in scientific activities. There is no broad consensus on any unified account of models, as stated by Gelfert (2017), and it is considered an obvious consequence of this void to assume that "if all scientific models have something in common, this is not their nature but their function" (Contessa 2010: 194). Moreover, if this characterization of models as functional entities is accepted, we must then specify how the models work as "carriers of scientific knowledge" (Ducheyne 2008: 120).

One of the basic relationships between the model and its target-system (T) that has to hold, if the model must carry scientific knowledge, is the representation.[1] My aim is not to advance a general theory of scientific representation, but

[1] See also Campbell 1920; Hesse 1966; Giere 1988; Morgan and Morrison 1999; Hughes 1997; Teller 2001; Van Fraassen 2008; and Mitchell 2013. Concerning the issue of scientific representations and realism, deeply tight, for a defense of scientific realism, see also Alai 2021a, 2021b, and 2023.

Argumenta 9, 1 (2023): 43-60
ISSN 2465-2334

© 2023 Giovanni Galli
DOI 10.14275/2465-2334/202317.gal

to propose a definition of the representational relationship between the specific kind of models produced by the deep-learning neural network system AlphaFold (AF), and their T. In §2 I present the main positions and definitions of models as functional entities. This, however, is mostly a study about the semantics of the representational relationship between AF and its T, for it is on the basis of that relation that such models are carriers of knowledge. Examples of this relationship regard models of actual T, such as the double-helix model of DNA, or the Bohr model of the atom, i.e. models that represent existing objects, and also models of potential (non-actual) T, as the examples of repressillators, synthetic oscillators and the ultra-Keynesian model analyzed by Knuuttila (2021), i.e. models that represent objects not existing in nature. According to the *representationalist* view what we learn from models presupposes a representational relation, while according to the *inferentialist* view, the representational feature of models is decoupled from their capacity of carrying knowledge. I claim that the representational relation presupposes the epistemic function of models of both actual and potential T. In §3 I discuss Knuuttila's (2021) artefactual view of models. In §4 I argue that the example of models of potential T does not invalidate the role of the representational relationship, and in §5 I discuss the contest of Critical Assessment of protein Structure Prediction (CASP) and AF. In §6 AlphaFold models are interpreted as simulation models. To conclude, in (§7) I argue that they hold a kind of morphic representational relation with their T. The general aim of this paper is to give one of the first contributions to expand a philosophical account of deep-learning models in general and AF models in particular.

2. A Taxonomy of Models

Models have a central role in sciences. Even if there is no consensus about their nature and qualifications, scholars have elaborated on three main areas: semantics, ontology, and epistemology of models. The first relates to what the models represent. The second concerns what the models are. The third focuses on the cognitive function modelers exploit for epistemological purposes. I will focus mainly on the first area, addressing namely the relation between the model and its target—system, specifically in the context of material, artifactual and simulation models, as they are tackled by Rosenblueth and Wiener (1945), Knuuttila (2021) and Durán (2018, 2020).

There are three main conceptions of the model–T relation: the similarity conception, i.e., models and their T are to some extent similar; the structuralist conception, i.e., models represent their T in virtue of a morphic relation between them; and the inferential conception, i.e., models as scientific representations have to be analyzed in terms of the inferential function.[2] Each conception offers different answers to certain problems. Moreover, we can distinguish the instantial view and the representational view. According to the former, models instantiate the axioms of a theory, that is composed of linguistic and mathematical statements. The representational view instead holds that it is rather the language that is connected with the model, while the model connects to the world "by way of similarity between a model and designated parts of the world" (Giere 1999: 56). In turn, the representational view has an informational and a pragmatic version.

[2] For a general discussion about the arguments and problems of the three accounts of scientific representations, see Frigg and Nguyen 2021.

The former conceives representation as "an objective relation between the model and its target, which imbues the former with information about the latter" (Gelfert 2017: 26). According to the latter, instead, it is not possible to "reduce the essentially intentional judgments of representation-users to facts about the source and target object or systems and their properties" (Suarez 2004: 768).

A further distinction can be drawn between substantive and deflationary accounts of representation. Substantive accounts aim for a robust explanation of the function of a representation in terms of a fundamental relation between a model and its target. Deflationary accounts, instead, settle for a light characterization of the functional unit of representational devices. We will see that while Knuuttila's proposal is pragmatic and deflationary, even though recognizes a representational function of models, the AF models are better interpreted by the representational, informational, and substantive view.

3. The Artifactual Account of Models

AF models, as representations of proteins, are a result of sophisticated techniques that make use of experimental data and abstract models. The 3d structures of proteins predicted by AF recall the structure of material models of a DNA strand but with a digital suit. One of the first studies on the representational capacity of models has been made by Wiener and Rosenblueth (1945). They analyze the role of material models of phenomena in scientific research, stressing their advantage with respect to abstract models thanks to their representational features. They describe a material model as "the representation of a complex system by a system which is assumed simpler and which is also assumed to have some properties similar to those selected for study in the original complex system" (Rosenblueth and Wiener 1945: 317). The relation identified by the authors between the material model and the original complex system can be seen as a case of similarity conception. This view then contrasts Suarez's inferential conception. These models are intended to be approximations and "surrogates" (Rosenblueth and Wiener 1945: 320) for the real facts under observation. But models can represent also facts not already present in reality. Indeed, Knuuttila is interested in developing an account of models consistent with the need, in some areas of inquiry as economics or synthetic biology, to build models of objects we do not find in nature or in society, i.e. models of invented objects.

Knuuttila (2021) advances the artifactual account of models which fits well with the inferential account developed by Suárez (2004). She is interested in stating an alternative position to the received ones, both substantive and deflationary, pointing out that models can be carriers of scientific knowledge even if they do not represent the actual state of affairs in the world. She insists on the modal reach (Godfrey-Smith 2006) and the modal dimension of modeling (Le Bihan 2016), "which approaches models as purposefully constructed systems of interdependencies designed to answer some pending scientific questions" (Knuuttila 2021: 5). Models as epistemic artifacts function as "erotetic devices" (Knuuttila 2021: 6). Such devices are artificial systems that deploy dependencies constrained to the aim of answering a specific scientific question, supported by theoretical, and empirical considerations.

Two examples are described, one of an ultra-Keynesian model as an example of an economic model that does not refer to a real T, and one of repressilators and synthetic oscillators in synthetic biology, that do not correspond to any existing

circuits, but are rather pictured to explore and test possible biological circuit designs. To strengthen the cases, she distinguishes between representational modes and media, and also between internal and external representations. The representational modes are the many semiotic devices that express various meanings and contents, while the representational media are for example the ink on paper, digital computer, biological substrata, and what support the representations. According to Knuuttila (2021: 5) the same representational mode can be implemented in different media as the example of the synthetic repressilator and the electronic repressilator that instantiate both the same ring oscillator design, yet they are implemented in different media "enabling different kinds of inferences" (Knuuttila 2021: 5). Moreover, an internal representation concerns "how various kinds of sign-vehicles or representational devices are used to make meaning and convey content" (Knuuttila 2021: 5), i.e. for a material model of the atom, the material, the proportion, and in general the semiotic and semantic features of the model chosen to represent the specific object; by external representation, instead, she refers "to the relationship of a model to a real-world target system, the question on which the philosophical discussion has largely concentrated" (Knuuttila 2021: 6). This distinction is particularly relevant for the definition of models as epistemic artifacts:

> Nevertheless, the fact that something may be internally represented within a model without necessarily representing the actual state of worldly affairs opens up the prospect of conceiving modeling as a practice of exploring the possible (Knuuttila 2021: 7).

The artifactual approach allows us to see biology as a discipline that not only focuses on natural organisms but includes also potential organisms (Elowitz and Lim, 2010, 889). So conceived, models are carriers of knowledge in virtue of their being erotetic devices and artifactual constructs useful to support surrogative inferences about a potential target-system. In such a way, inferentialists would argue that their representational capacity is not relevant to their use in exploring the possible.

4. Some Remarks on the Artifactual Account of Models

The artifactual account stresses the pragmatic goal that directs the models' construction and manipulation. It is to conceive models as tools for investigating specific phenomena, used to answer scientific questions, motivated by theoretical, and empirical tenets. According to Knuuttila (2021), their accomplishment relies on their modal function of exploring the spaces of possibilities and the main point is that their success needs not be grounded on the representational relation between the model and the target system. Thanks to the distinction between internal and external representations, Knuuttila safeguards a slightly deflationary definition of representation, which connects the artifactual models with a possible organism. Obviously, the correctness of models of merely possible T does not need the same kind of warrants as the models of real T. What does then warrant them? For Knuuttila it is simply their predictive success, without any need to invoke to any representational relation, yet it remains unanswered the question concerning what warrants the models' success. In other words, how can we probe the success

of a model of a potential target-system, without any reference to the representational relation between the model and the possible state of affairs? Knuuttila (2021) claims that it is still sufficient for a modal relation to justify the success of the artifactual models.

We can reframe the modal feature of the relation between the models and the potential T as a predictive relation, i.e., a model would predict the possible state of affairs, if there were conditions such and such. One of the kinds of models so far used to explore the possible phenomena within a manifold scenario is the simulation model (SM). That is a model resulting from computational procedures able to predict or determine specific output with a given set of data. SM are helpful to study and predict complex scenarios and phenomena. They are implemented by a certain degree of idealization and can be used to study *actual* T (like biological systems, i.e. birds flocks, ant colonies, structure determination, enzyme kinetics and molecular dynamics) and *potential* T (like the behavior of mechanics and artifacts as airplanes, spacecrafts, biomedical robots, and also new proteins, new drugs and possible organisms). As it happens with imaginary economics, repressilators and oscillators, from a set of data and techniques the respective models predict how the possible systems would act. To this extent, artifactual models are a kind of simulation model: though the examples are not strictly speaking computer-based simulations, they simulate possible states of affairs, useful to predict how the system will work.

I submit, however, that neither for simulation models nor for material models we can easily dismiss the representational link between the model and its T. In the case of artifactual models, it seems intuitive not to stress the representational link, because we weigh differently the conceptual role of an actual T and a potential T. However, if we want to gain epistemic access to the T in question, actual or potential, the model has to maintain a representational link with it. I call it the accessibility condition (AC):

> *Accessibility condition*: A model M of a target-system T is a functional carrier of knowledge in virtue of its capacity to give epistemic access to T through the representational relation established by the researchers between M and T.

In the case of AF the output models of predicted proteins' structures can be conceived as a kind of artefactual model. Most AF models represent actual target systems, but they are also useful in the exploration of potential proteins. In that case, their success depends on their accurately representing the modal properties of proteins, i.e., what is actually possible or impossible for proteins. The discussion on representation, then, is far from over, and a substantive view of representation is still in play.

5. CASP and AlphaFold Protein Structure Prediction

AF is a breakthrough deep-learning network AI system able to predict highly accurate protein structures.[3] Its computational power and sophisticated engineering let the DeepMind team, which worked on it, win the CASP 14 (Critical Assessment of Protein Structure Prediction) on the 30th of November 2020. CASP started in 1994 and it is a biennial competitive appointment for biological researchers working on protein structure prediction, aiming at solving the well-

[3] All the predicted structure can be found on the AF open access database here: https://alphafold.ebi.ac.uk/ (last access November 2023).

known folding problem: How is it possible to fold a protein starting from its strains of amino acids? The founder and chair of CASP is John Moult, Professor of the Institute for Bioscience and Biotechnology Research and the Department of Cell Biology and Molecular Genetics at the University of Maryland. He describes CASP in this way:

> Computational biology differs from traditional science in that it takes place in a virtual world. Achieving rigor in a computational world which the scientist controls is much harder than when dealing with the inflexible realities of the physical world. We introduced Community assessment experiments in computational biology to help achieve the same rigor as in real world science. CASP (Critical Assessment of Structure Prediction), the first framework for these experiments, is an organization that conducts double blind community wide experiments to determine the state of the art of computational methods for modeling protein structure from amino acid sequence and other information. CASP has now been running for over 20 years, with continuing high participation rates (over 100 groups around the world), and has been accompanied by an enormous improvement in the accuracy of the protein modeling methods. The CASP methodology has now been adopted in a wide range of computational biology areas, including protein-protein interactions, genome sequence annotation, biological networks, and protein function annotation (Moult 2022).

The first lines make a sharp distinction between the rigor achieved in the real-world sciences and the one obtained in a computational world. I am interested in showing the philosophical relevance of the effort to make the two methodologies meet and enhance each other. Two questions. Why do the real-world sciences working on protein folding need such an upgrade? Moreover, why is it so important to solve the folding problem? "We have discovered more about the world than any other civilization before us. But we have been stuck on this one problem. How the proteins fold up. How the protein goes from a string of amino acids to a compact shape that acts as a machine and drives life?",[4] says John Moult (2021), filmed in *AlphaFold: The making of a scientific breakthrough*, the inside story of DeepMind[5] research team who created AF. This is indeed the folding problem. Solving it means making huge steps in molecular biology and consequently in many other biological fields. DeepMind team states that the research program that leads to AF and similar systems is crucial for the development of the life sciences. Proteins are stunning biological nano-machines, whose understanding will take us to unveil how they work and interact with other molecules. They are polymers in which the 20 natural amino acids are connected by amino bonds. They are polymers in which the 20 natural amino acids are connected by amino bonds. They are synthesized by the ribosomes, which are complex molecular machines present in all living cells, measuring around 30 nm. Ribosomes compose amino acids together in the specific order defined by messenger RNA molecules.

[4] John Moult was interviewed in *AlphaFold: The making of a scientific breakthrough*, video interview about the AlphaFold breakthrough: https://www.youtube.com/watch?v=gg7WjuFs8F4 (last access November 2023).

[5] AlphaFold thematic section on DeepMind website: https://www.deepmind.com/research/highlighted-research/alphafold (last access November 2023).

AF team trained this system[6] on publicly available data consisting of around 170.000 protein structures taken from the protein data bank (PDB),[7] together with large databases containing protein sequences of unknown structure. Thanks to the genomics revolution we can read amino acid sequences of proteins at massive scale; in fact, the Universal Protein database (UniProt) contains 180 million protein sequences. The building blocks of proteins are amino acids, small molecular compounds with unique features composed of between 10 and 20 atoms. In ordinary biology we find 20 standard types of amino acids floating within the cytoplasm of the cells. They connect to a piece of transfer RNA that matches with the three genetic sequences of the genetic code of the RNA messenger. Ribosomes then read the three-basis instructions of the RNA messenger and start building a chain of amino acids that goes out from the ribosome. As the chain of amino acids exits the ribosome, released in the cytoplasm, it is surrounded by water molecules and subject to the interaction of physical forces that make the chain fold up on itself and form the complex 3d structure we call a protein. All this process is called translation because the molecular mechanisms manage to produce a fully operative protein with proper functions by translating a piece of the genetic code. The unique shape of a protein is defined by its amino acid sequence and its shape is the key to unlock its functions. Determining the 3d structure of a protein is indeed necessary to understand its functions. Proteins seem like pieces of a puzzle, but with a dynamic shape which can change according to the bonds they make with other interacting molecules. Nonetheless, a protein would bond with some molecules and not others. There are specific combinations of proteins and molecules. By understanding the protein shape and the occurring molecular interactions, scientists can design vaccines, new drugs and functional structures for ecological purposes: "Among the undetermined proteins may be some with new and exciting functions and—just as a telescope helps us see deeper into the unknown universe—techniques like AlphaFold may help us find them" (The AlphaFold Team 2020).

Proteins are fundamental for most living beings, and enhancing their understanding through computational allows us to tackle diseases, discover new medicines and disclose the enigmas of life in a faster and cheaper way than traditional research on existing proteins. Thanks to painstaking experimental effort, real-world sciences have determined before the release of AF the 3d structures of approximately 100.000 unique proteins (Thompson, Yeates and Rodriguez 2020; Bai, McMullan and Scheres 2015; Jaskolski, Dauter and Wlodawer 2014; Wüthrich 2001). Using the experimental methodology scientists had at their disposal until now, it could take from months to years and a lot of financial resources to determine a single protein structure. Computational methodologies are in fact needed to reduce this gap and to "enable large-scale structural bioinformatics" (Jumper, Evans, Pritzel et al. 2021a: 1). That is why CASP has been promoted within the biological fields, with the aim to push researcher communities to solve the protein folding problem, that has been an open research problem since when,

[6] It makes use of 16 TPUv3s (which is 128 TPUv3 cores or roughly equivalent to ~100-200 GPUs) run over a few weeks, a relatively modest amount of compute in the context of most large state-of-the-art models used in machine learning today. See Jumper, Evans, Pritzel et al. 2021a.

[7] Protein Data Bank website: https://pdb101.rcsb.org/ (last access November 2023).

around 1960, the first atomic-resolution protein structures were proposed (Kendrew 1961; Pauling and Corey 1951; Pauling, Corey and Branson 1951), while the first protein structures detected presented unpredicted irregularities. It was the case of globin structures, a clade of globular proteins containing heme, a precursor to hemoglobin (6,5 nm), involved in binding, and transporting oxygen. Globin proteins contain the globin fold, which is a series of eight α-helices packed together in irregular ways. Since the 60's the folding problem concerns three different problems (Dill, Ozkan, Shell and Weikl 2008):

1) The folding code: the thermodynamic question of what balance of interatomic forces dictates the structure of the protein, for a given amino acid sequence;
2) Protein structure prediction: the computational problem of how to predict a protein's native structure from its amino acid sequence;
3) The folding process: the kinetics question of what routes or pathways some proteins use to fold so quickly. We focus here only on soluble proteins and not on fibrous or membrane proteins.

The main CASP evaluation follows the criteria of comparison between the predicted model α-carbon positions and those in the real-world target structure. The visualisation of cumulative plots of distances between pairs of α-carbon in the model and target structure positioning is used to evaluate the prediction against the experimental result, such as shown in the two figures aligning computational prediction with the experimental result. The real structure is already known by the evaluator so that the CASP examination can estimate the accuracy of the predictive model. To each prediction is assigned a numerical score GDT-TS (Global Distance Test—Total Score) specifying the percentage of modeling residues[8] in the model with respect to the target.

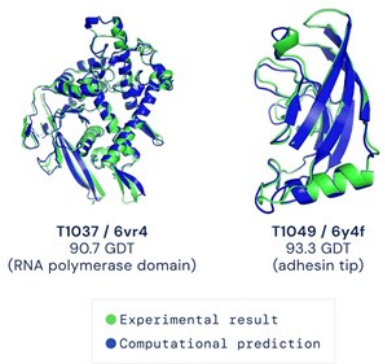


Figure 1: Two examples of protein targets in the free modelling category. AlphaFold predicts highly accurate structures measured against experimental result (The AlphaFold Team 2020).

The CASP campaign evaluation relies basically on the issues of 1) The folding code, 2) Protein structure prediction, and 3) The folding process, although the

[8] The amino acids in a polypeptide chain are linked by peptide bonds. Once linked in the protein chain, an individual amino acid is called a residue, and the linked series of carbon, nitrogen, and oxygen atoms are known as the main chain or protein backbone.

results are carried out in many prediction categories: tertiary structure prediction, residue-residue contact prediction, disordered regions prediction, function prediction, model quality assessment, model refinement, and high-accuracy template-based prediction. Tertiary structure prediction is then divided into three subcategories: homology modeling; fold recognition; and *de novo* structure prediction (New Fold). All these conditions form what we can call the *accuracy qualification* (AQ). The higher the GDT scores, the better the AQ of the predictions, and the higher the AQ, the nearer the model to the real shape of the protein. Another consequence of the AQ is that higher scores correspond to higher amounts of correct information transmitted from T to M, and from M to the modelers.

Since 2018 CASP team made some improvements, but the big leap was between AlphaFold 1 (AF1), the ancestor, and its successor, AlphaFold 2 (AF2), whose score, according to Moult, was around 90 GDT on 100 points scale prediction accuracy. DeepMind developed new deep learning architectures to improve the research methods for CASP14, which led to a high level of accuracy. These methods are inspired by the research areas of biology, physics, and machine learning and by the studies many scientists enhanced during the years on the protein folding problem. The AF2 system is described as a neural network-based model (Jumper, Evans, Pritzel et al. 2021a). It is important to note that it is described as an AI system coherent with the wider project of Demis Hassabis, CEO and co-founder of DeepMind, of making further steps in General AI. The whole AF architecture learns from the data and elaborates the 3d structure prediction of the folded protein. We can think of a folded protein as a spatial graph, a spatial presentation of a graph in the 3-dimensional Euclidean space R^3, in which residues are the nodes and edges link the closely related residues (Jumper, Evans, Pritzel et al. 2021a). The graph matters to understand the proteins physical interactions and their evolution. For the second version of AF2, the team created an attention-based neural network system, trained end-to-end, that attempts to interpret the structure of this graph while reasoning over the implicit graph that it's building (Jumper, Evans, Pritzel et al. 2021a). By process iteration, AF2 produces accurate predictions of the underlying physical structure of the protein in daystime. Moreover, the system can predict the reliability of parts of each predicted protein structure using an internal confidence measure. The following is the AF1 architecture that provided important results in CASP13, beating the median free-modeling accuracy of other systems.

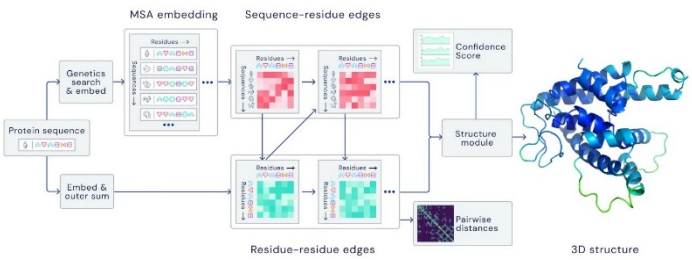


Fig. 2: An overview of the main neural network model architecture. The model operates over evolutionarily related protein sequences as well as amino acid residue pairs, iteratively passing information between both representations to generate a structure (The AlphaFold Team 2020).

AF1 has a straightforward architecture (Senior, Evans, Jumper et al. 2020). It begins with the amino acids sequence for which we are searching the protein structure. The first step concerns a data extraction move from the known database, in order to find similar protein sequences. The first task of the neural network is to find similar sequences, and it is called Multiple Sequence Alignment (MSA). The protein structure is responsible for its function, and we know that evolution carved the organisms in such a way that only some structures passed the survival threshold. Indeed, in different organisms during evolution a protein structure is more stable over time than the genetic sequence encoding that particular protein the genetic mutations that passed the evolutionary test are those that did not affect the protein structures. Comparing evolutionary-related protein sequences, whose 3d form should share some similarities, is what MSA does: scrolling the database to find amino acid sequence matches in the animal kingdom. To sum up, in AF1, 3 main steps need to aim at structure prediction:

1) AF1 collects the MSA features;
2) it predicts then the distogram using a residual neural-network;
3) it optimizes the protein backbone using the predicted distogram in combination with simulated physical forces. The output is the 3d predicted protein structure.

As the aforementioned system, AF2 presents three main blocks:

1) A pre-processing stage where the input sequence is used to query additional information about the initial sequence from databases;
2) The information is then mapped into an MSA and pair representation, which are refined by the Evoformer, a 48-layer deep transformer-like network that uses attention mechanisms to update MSA and pair representations;
3) The structure module, a recurrent network, processes the Evoformer output, which transforms the abstract representations of the Evoformer into concrete 3d coordinates of the protein geometry.

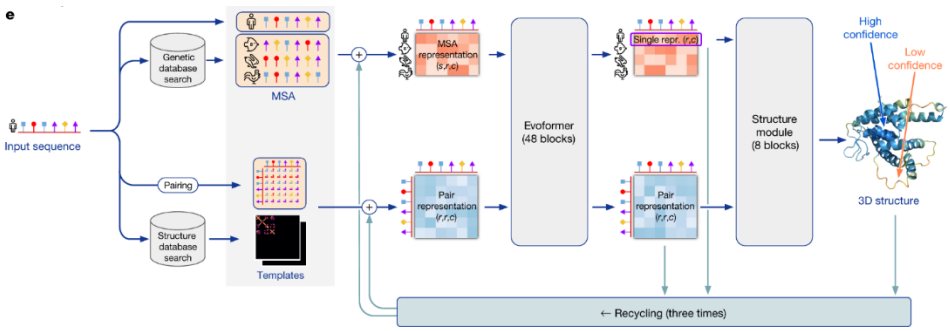


Fig. 3: Model architecture. Arrows show the information flow among the various components described in this paper. Array shapes are shown in parentheses with s, number of sequences (Nseq in the main text); r, number of residues (Nres in the main text); c, number of channels (Jumper, Evans, Pritzel et al. 2021a).

Just to cite the improvement of the new architecture of AF2, it allows for jointly embedding of multiple sequence alignments (MSAs) and pairwise features. Moreover, AF2 has a new representation output and an associated loss that together allow for end-to-structure prediction.
The AF research teams does not submit that the system is capable of revealing underlying laws regulating protein folding. AF, however, seems to have reached important results concerning some kinds of proteins, especially those based on a strain of between 100 and 200 amino acids. Moreover, albeit the neural networks system distances the empirical link of evidence gathered from experimental data in the genomic database of proteins, it has the computational power to disclose the structure of the simulated object. In future, it may be capable of finding common patterns between the structures predicted. In any case, from a philosophical perspective, it is important to ask whether this kind of AI system can assist researchers in unveiling recurrent structures that could be defined as the laws governing protein folding. This discovery could improve even better the system solving the folding problem.

6. AlphaFold as a Simulation Model

In the last years, as the use of deep-learning neural networks has become pervasive in engineering and scientific areas,[9] scholars have focused correspondingly on the diffusion of simulation models as tools and outputs of neural network systems. What are simulation models[10] is then a crucial issue in the epistemology of models and the general philosophy of sciences.

A simulation model (SM) is a representation of a real or possible system, interacting with a determined environment, supported by computation techniques and expressed through visualization tools. It is a powerful instrument to represent, observe, study and manipulate to a higher degree of realism complex phenomena within a system. I submit that a model produced by AF is a kind of SM endowed with a degree of accuracy that was not available in the past, therefore improving the representational link between M and the related T. I submit that AF is a system architecture that produces SM of proteins' structures. We can divide AlphaFold into three main sectors: 1) AF as a complex neural-network system as a whole architecture; 2) AF sector sequences of algorithmic processing, the main blocks of the architecture; 3) AF's protein structure model as the output of the system.

As we know, the first stages of the system have to do with the analysis of the protein structure data contained in the database. In fact, in CASP the accuracy of the predicted structure is measured through the structure model obtained via experimental methods through X-ray crystallography and NMR spectroscopy. I claim that in each sector AF works as a kind of SM. According to CASP14 there are three relations to be noted:

1) The first between the real target system T and the experimental model, i.e. the relation between the real receptor-binding protein adhesin (Fig. 1) and the model resulting from the use of X-ray crystallography and NMR spectroscopy;

[9] See also Mitchell 2019; and Wooldridge 2021.
[10] See also Durán 2018, 2020; and Paronitti 2008, 2009.

2) The second between the experimental model and the simulation model AF, namely what is pictured in Fig1, the relation between the model obtained through experimental methods and the simulation model produced by AF;

3) The third between AF, the whole system architecture and the real target-system T, i.e. the real adhesin protein. The experimental and simulation success of these models is due to the relation they have with T. In the first case, the relation is obtained through experimental work which preserves the empirical link between observation and data manipulation. In the second case, the two different kinds of models are both successful representations of T, even though the simulation success entails a higher abstraction than the empirical link the experimental model holds in the first place. In the third case, AF as a whole architecture and the target system are not linked by an empirical relation, in so far as there is no direct observational contact as in the case of X-ray crystallography or NMR spectroscopy between the enquirer and the T. They are connected through the data manipulation and the simulation process binding the initial data, with the structure model in output.

Given the digital, computational, and algorithmic nature of the AF system, we can interpret it as an architecture producing simulation model (SM). There are mainly two types of simulation models: 1) SM is conceived as an implementation of models already existing; for example, aerospace engineers use SM of planes to test models they already have under specific circumstances like mechanical stress and weather conditions; 2) CS as models which have their own complexity and autonomy, the study of which is enhanced focusing on computer science and software engineering.[11] According to Durán (2021: 317), a simulation model (SM) is a "rich and complex structure that departs in important ways from standard models used in scientific research". Furthermore, Durán (2021) argues that the construction of the SM is possible because of a new methodology that is in place. He calls it *recasting*, and it consists of clustering a multiplicity of models into one fully computational SM. Think of it as the mashing-up of different models, also theoretical and mathematical, that could be implemented through deep-learning networks, with the specific aim to predict, in this case, the folding of proteins. To refine the terminology for the purposes of AF, we can call the methodology in place *reshaping*. AF begins with a set of data with empirical and experimental information, then through the intervention of programmers in adjusting the learning bias with respect to the desired output, using different integration modules, idealisations, and reshaping the data representation with the multiple sequence alignments MSA, according to cycles of implementation and integration, through the Evoformer and the Structure Module, we gain the visualization of the 3d geometry of the folding shape of the protein.

Not all the SM produced by AF are accurate representations of their T, especially the complex proteins are very hard to predict through the AF architecture as it is. Moreover, AF does not predict important aspects of protein structures as many ligands, metal ions and cofactors. Furthermore, the main limitation of AF is that the system predicts only a single state of the protein, and it is also hard to tell which state of the protein will be represented by the model (Perrakis and Sixma 2021). In fact, AF produces indeed SM with specific aims and empirical and theoretical assumptions and limitations, that must pass the abovementioned *accessibility condition* AC. Moreover, given the accuracy standard gained from the

[11] See also Symons and Alvarado 2019; Durán 2018; and Boyer-Kassem 2014.

experimental data, we can draw another requirement to be satisfied, the correctness condition (CC) for the proteins models:

> *Correctness condition*: SM represents correctly iff the accuracy qualification (AQ) is satisfied.

The AQ developed by CASP is a threshold for the correctness of the representation of SM. I take it as the level of approximation to reality the representation gains from the system through the work of modelers.

To conclude, AF consists not only of a complex and sophisticated computational implementation of the experimental models of proteins' structure determination, but it is a simulation model which is already changing the scenario of the computational and structure biology research areas.

7. Structure and Representation

I have advanced an interpretation of AF models as simulations. Thanks to the simulation power, modelers have greatly improved the representational capacity of models. Now I suggest a definition of the relation, refined through simulation, holding between the AF models and the objects they aim to represent:

> *Structural Dynamic Approximate Isomorphism*: a mapping that gathers through simulation even more information about the dynamic structure of T, so that the two systems (the model and T) approximately share the main structural features.

This definition pictures the ideal isomorphism between the model and the real protein which AF assumes as an implicit presupposition. It is a form of mapping since AF aims to visualize the shape of the protein as an image which can be navigated and observed in many aspects on a computer. The two systems should share the same features, represented one-to-one in the model: the individual folding units (domains), dynamic movements, contact matrix, ligands, and each polypeptide chain, and monomers, involved in multimers. Moreover, the two systems should share the same features under the same dynamics, i.e. the interactions of the domains in T should correspond in the mapping of the model. Given the limitation of AF, the definition assumes that the simulation model could be refined through time thanks to more and better information about the relevant features of the real proteins. The isomorphism between the two systems should regard the geometry as the information detected regarding the ligands and the folding units. In the case of protein folding the isomorphic relation is fundamental between the two systems, in so far as the protein shape is responsible for its function.

Why should the isomorphism be dynamic? One of the most important limitations of AF is that it predicts only a single state of a protein, but the aim of the AF researcher is to overcome this boundary. AF models are the peak of an important history of views about, and scientific representation of, proteins. In the last century structure biologists[12] shifted from the static view, according to which the protein models represented rigid structures, to the dynamic view:

[12] For a review of the history of structural biology, from the static to dynamic view, and a philosophical account of representation and explanation in the study of protein in structural biology, see Neal (2021).

> The study of how proteins serve the needs of a living organism is a curious case in which a method that yielded dramatic advances also led to a misconception. The method is X-ray crystallography […] The intrinsic beauty and the remarkable detail of the structures obtained from X-ray crystallography resulted in the view that proteins are rigid. This created the misconception, namely that the atoms in a protein are fixed in position (Karplus and McCammon 1986: 42).

The dynamic turn in protein representations owes a lot to thermodynamics. In fact, the dynamic analysis treats proteins as thermodynamic systems. The shift brought changes also to the structural concept. The old structural concept, coherent with the static view, is committed to the beliefs 1) that every protein has a rigid and static 3d structure and 2) that the protein structure alone determines protein function. The new dynamic concept of protein structure drops these commitments and adopts an inferential stance toward the proteins' structures, which are taken to be flexible, dynamic and constantly under structural fluctuations and mutations according to the environment and occurrent phenomena. Advocates of the dynamic concept are committed to the belief that dynamics and structures are relevant determinants of protein behavior and function (Neal 2021). The supporters of the dynamic concept suggest a wide range of experimental, theoretical and computational strategies to test the dynamic properties of proteins. AlphaFold researchers support the dynamic view of protein structure, well represented by accurate prediction models.

The motivation of AF is that biological research will be aided by the availability of an open-source determination structure database. The assumption underlying AF system and fostering this motivation is that simulation model structures entail an isomorphic relation with the target-protein. The protein may be in the real world, or a possible protein, or a protein mutation, whose structure is to be explored, in order to accomplish some specific functions, as in the case of PET depolymerization (Lu, Diaz, Czarnecki et al. 2022). AF model assumes that the dynamic view can be fostered through computational methods via deep-learning network architecture.

The AF system architecture is built to replicate the shape of the proteins according to their geometric features. The SM is apt to replace the representation of a protein given by the experimental procedures. The accuracy of the AF models is then grounded on the approximation to the structure of the real protein or to the functional structure of potential proteins. What best captures the conservation of information and geometric features between M and T is the notion of isomorphism. Related to protein structure prediction or drug discovery, AF researchers are therefore committed to a kind of isomorphism. On its basis, we can then define the representation relation:

> *Representation*: A scientific model M represents a T, which may be actual or potential, iff the dynamic structure of the model is approximately isomorphic to the structure of the T.

This kind of definition avoids some problems described in the structuralist conception of scientific representations.[13] According to Suárez (2003) and Downes (2009) isomorphism cannot ground the representation relation, because the former is characterized as reflexive and symmetrical, while the latter is not. Frigg and Nguyen (2017: 55) coin the requirement of directionality to account for this asymmetry. To

[13] See also Gelfert 2017; Frigg and Nguyen 2021.

answer these critics, let us recall that AF modelers do not aim at ideal models of proteins. The 100% GDT score is an ideal limit of research output, while the condition to be obtained is the standard of accuracy, i.e. AF models are accurate in so far as they represent their T, as an experimental representation of them would have done. The accuracy of AF models relies on the training the networks have got from the experimental data gathered. The isomorphic relation is approximate in the sense that the relation safeguards the correctness condition (CC).

Moreover, since the function of a protein depends on its folding, in the dynamics of interaction with the phenomena and molecules in the environment, there is a fundamental connection between the information it carries and the structure it takes once folded. Modeling such a dynamic structure allows us to understand the function of the protein. The isomorphism between the target-structure and the simulated or predicted structure is crucial to study, manipulate, and explore actual and possible functions of proteins. In so far as we need models to offer information about the target, the directionality of representation is then from model to target. It is indeed the asymmetry of the M-T relation that assures the accessibility condition (AC) that accurate models accommodate.

The isomorphic picture of the representational relation between the AF models and their T is one to take at face value if we want to develop a philosophical account of a breakthrough scientific advance such as AlphaFold.[14]

References

Alai, M. 2021a, "Scientific Realism and Further Underdetermination Challenges", *Axiomathes*, 31, 779-89.

Alai, M. 2021b, "The Historical Challenge to Realism and Essential Deployment", in Lyons, T.D. and Vickers, P. (eds.), *Contemporary Scientific Realism: The Challenge from the History of Science*, Oxford: Oxford University Press, 183-215.

Alai, M. 2023, "Scientific Realism, Metaphysical Antirealism and the No Miracle Arguments", *Found Sci*, 28, 377-400.

Bai, X.C., McMullan, G., and Scheres, S.H.W. 2015, "How Cryo-EM Is Revolutionizing Structural Biology", *Trends Biochem. Sci.*, 40, 49-57.

Boyer-Kassem, T. 2014, "Layers of Models in Computer Simulations", *International Studies in the Philosophy of Science*, 28, 4, 417-36.

Campbell, N.R. 1920, *Physics: The Elements*, 1st edition, Cambridge: Cambridge University Press.

Contessa, G. 2010, "Editorial Introduction to Special Issue", *Synthese*, 172, 2, 193-95.

Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R. 2008, "The Protein Folding Problem", *Annu Rev Biophys.*, 37, 289-316.

Downes, S. 2009, "Models, Pictures, and Unified Accounts of Representation: Lesson from Aesthetics for Philosophy of Science", *Perspective on Science*, 17, 4, 417-28.

Ducheyne, S. 2008, "Towards an Ontology of Scientific Models", *Metaphysica*, 9, 1, 119-27.

[14] I would like to thank Mario Alai and the two anonymous referees for their helpful comments along with the participants in the PRIN conference "The Manifest Image and The Scientific Image" held in Urbino, on the 20th and 21st of June 2022.

Durán, J.M. 2018, *Computer Simulations in Science and Engineering: Concepts—Practices—Perspectives*, Berlin: Springer.

Durán, J.M. 2020, "What Is a Simulation Model?", *Minds & Machines*, 30, 301-23.

Elowitz, M.B. and Lim, W.A. 2010, "Build Life to Understand It", *Nature*, 468, 7326, 889-90.

Fischer, E. 1894, "Einfluss der Konfiguration auf die Wirkung der Enzyme", *Berichte der deutschen Chemischen Gesellschaft*, 27, 3, 2985-93.

Fischer, E. 1906, *Untersuchungen über Aminosäuren, Polypeptide und Proteine*, Vol. 2, Berlin: Springer.

Frigg, R. and Nguyen, J. 2017, "Models and Representation", in Magnani, L. and Bertolotti, T. (eds.), *Springer Handbook of Model-Based Science*, Dordrecht: Springer, 49-102.

Frigg, R. and Nguyen, J. 2022, "Scientific Representation", *The Stanford Encyclopedia of Philosophy*, Zalta, E.N. (ed.), https://plato.stanford.edu/archives/win2021/entries/scientific—representation/

Gelfert, A. 2016*, How to Do Science with Models: A Philosophical Primer*, Berlin: Springer.

Giere, R. 1988, *Explaining Science*, Chicago: University of Chicago Press.

Giere, R. 1999, "Using Models to Represent Reality", in Magnani, L., Nersessian, N.J., and Thagard, P. (eds), *Model-Based Reasoning in Scientific Discovery*, New York: Plenum Publishers, 41-57.

Godfrey-Smith, P. 2006, "The Strategy of Model-Based Science", *Biology and Philosophy*, 21, 725-40.

Hesse, M.B. 1966, *Models and Analogies in Science*, Notre Dame: University of Notre Dame Press.

Hughes, R.I.G. 1997, "Models and Representation", *Philosophy of Science*, 64, 325-36.

Jaskolski, M., Dauter, Z., and Wlodawer, A. 2014, "A Brief History of Macromolecular Crystallography, Illustrated by a Family Tree and Its Nobel Fruits", *FEBS J.*, 281, 3985-4009.

Jumper, J., Evans, R., Pritzel, A. et al. 2021a, "Highly Accurate Protein Structure Prediction with AlphaFold", *Nature*, 596, 583-89.

Jumper, J., Evans, R., Pritzel, A. et al. 2021b, "Supplementary Information for Highly Accurate Protein Structure Prediction with AlphaFold", *Nature*, 596, 583-89.

Karplus, M. and McCammon J.A. 1986, "The Dynamics of Proteins", *Scientific American*, 254, 4, 42-51.

Kendrew, J.C. 1961, "The Three-Dimensional Structure of a Protein Molecule", *Scientific American*, 205, 6, 96-111.

Kendrew, J.C., Bodo, G., Dintzis, H. et al. 1958, "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis", *Nature*, 181, 662-66.

Knuuttila, T. 2021, "Epistemic Artifacts and the Modal Dimension of Modeling", *Euro Jnl Phil Sci*, 11, 65.

Le Bihan, S. 2016, "Enlightening Falsehoods: A Modal View of Scientific Understanding", in Grimm, S.R., Baumberger, C., and Ammon, S. (eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, London: Routledge, 111-35.

Lu, H., Diaz, D.J., Czarnecki, N.J. et al. 2022, "Machine Learning—Aided Engineering of Hydrolases for PET Depolymerization", *Nature*, 604, 662-67.

Mitchell, S. 2013, *Unsimple Truths: Science, Complexity, and Policy*, Chicago: University of Chicago Press.

Mitchell, M. 2019, *Artificial Intelligence: A Guide for Thinking Humans*, New York: Farrar, Straus and Giroux.

Moult, J. 2020, "AlphaFold: The Making of a Scientific Breakthrough", video interview, https://www.youtube.com/watch?v=gg7WjuFs8F4 (last access November 2023).

Moult, J. 2022, "CASP, Research Project", http://moult.ibbr.umd.edu/projects.htm (last access November 2023).

Morgan, M.S. and Morrison, M. 1999, "Models as Mediating Instruments", in Morgan, M.S. and Morrison, M. (eds), *Models as Mediators*, Cambridge: Cambridge University Press, 10-37.

Morrison, M. 2008, "Models as Representational Structures", in *Nancy Cartwright's Philosophy of Science*, Hartmann, S., Hoefer, C., and Bovens L. (eds.), Abingdon: Routledge, 67-88.

Neal, J.P. 2021, "Protein Structure, Dynamics and Function: A Philosophical Account of Representation and Explanation in Structural Biology", Doctoral Dissertation defended on July 27, 2021, https://d—scholarship.pitt.edu/41715/13/NealJP%20ETD%20History%20%26%20Philosophy%20of%20Proteins.pdf (last access November 2023).

Paronitti, G. 2008, *Epistemologia della simulazione: L'artificiale tra astrazione e realtà*, Morrisville: Lulu Press.

Paronitti, G. 2009, *Che cos'è la simulazione*, Roma: Carocci.

Pauling, L. and Corey R.B. 1951, "Atomic Coordinates and Structure Factors for two Helical Configurations of Polypeptide Chains", *Proc Natl Acad Sci USA*, 37, 235-40.

Pauling, L., Corey, R.B., and Branson, H.R. 1951, "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain", *Proceedings of the National Academy of Sciences*, 37, 4, 205-11.

Perrakis, A. and Sixma, K.T. 2021, "AI Revolutions in Biology: The Joys and Perils of AlphaFold", *EMBO Reports*, 22, 11, 1-6.

Perutz, M.F. 1970, "Stereochemistry of Cooperative Effects in Haemoglobin: Haem-Haem Interaction and the Problem of Allostery", *Nature*, 228, 726-34.

Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., and North, A.C.T. 1960, "Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å: Resolution, Obtained by X-Ray Analysis", *Nature*, 185, 416-22.

Rosenblueth, A. and Wiener, N. 1945, "The Role of Models in Science", *Philosophy of Science*, 12, 4, 316-21.

Senior, A.W., Evans, R., Jumper, J. et al. 2020, "Improved Protein Structure Prediction Using Potentials from Deep Learning", *Nature*, 577, 706-10.

Suárez, M. 2003, "Scientific Representation: Against Similarity and Isomorphism", *International Studies in Philosophy of Science*, 17, 225-44.

Suárez, M. 2004, "An Inferential Conception of Scientific Representation", *Philos. Sci.*, 71, 5, 767-79.

Symons, J. and Alvarado, R. 2019, "Epistemic Entitlements and the Practice of Computer Simulation", *Minds and Machines*, 48, 4, 729.

Teller, P. 2001, "Twilight of the Perfect Model", *Erkenntnis*, 55, 393-415.

The AlphaFold Team 2020, "AlphaFold: A Solution to a 50-Year-Old Grand Challenge in Biology", https://www.deepmind.com/blog/alphafold—a—solution—

to—a—50—year—old—grand—challenge—in—biology (last access November 2023).

Thompson, M.C., Yeates, T.O., and Rodriguez, J.A. 2020, "Advances in Methods for Atomic Resolution Macromolecular Structure Determination", *F1000Res.*, 9, 667, 1-18.

Van Fraaseen, B. 2008, *Scientific Representation*, Oxford: Oxford University Press.

Weisberg, M. 2007, "Who Is a Modeler?", *The British Journal for the Philosophy of Science*, 58, 2, 207-33.

Wooldridge, M. 2021, *The Road to Conscious Machines: The Story of AI*, London: Pelican.

Wüthrich, K. 2001, "The Way to NMR Structures of Proteins", *Nat. Struct. Biol.*, 8, 923-25.

Fiction and Reality: An Uncanny Relationship

Lisa Zorzato

University of Tartu

Abstract

In this paper, I will deal with the use of fictional models in the context of the realism *vs* antirealism debate. Specifically, I will argue that the explanatory role of fictional models can be accommodated by scientific realism. I will refer to the work of Alisa Bokulich, who has proposed a modification of realism in order to account for explanations employing fictional models. My own approach will be to offer an alternative: instead of a modification of realism, I will propose a modified notion of representation. Based on the work of James Clerk Maxwell and Bokulich's own account of it, I will introduce the notion of a 'ladder of abstractions', meaning an hierarchical organisation of mathematical structures constituting both models and theories. In this way, fictional model explanations can be construed realistically if understood as offering partial representations of a physical situation corresponding to an appropriate level of abstraction.

Keywords: Models, Fiction, Representation, Explanation.

1. Introduction

The term 'fictional models' will signify in the following "theoretical structures describing physical systems that are not, in fact, instantiated" (Zorzato 2023).[1] Fictional models are often considered to be problematic in terms of the debate between scientific realism and anti-realism. It would appear that their role is confined to being merely tools for calculations and predictions. However, fictional models can contribute positively to scientific explanation. Alisa Bokulich, in a book and a series of papers, has offered plenty of cases demonstrating that fiction can be 'a vehicle for truth' (Bokulich 2016). My main concern in this paper is to see how the use of fictional models can be accounted for from a realist viewpoint. In general, I agree with Bokulich when she says that her "account of explanatory fictions lies within a broadly realist approach to science" (Bokulich, 2016: 261).

[1] It is important to stress that the interest here is not in the ontology of scientific models; the question whether models are fictions or abstract entities will not concern me (for such questions see, e.g., Hendry and Psillos 2007, and Fiora Salis 2019).

Argumenta 9, 1 (2023): 61-72
ISSN 2465-2334

© 2023 Lisa Zorzato
DOI 10.14275/2465-2334/202317.zor

However, I shall argue that the question of how to relate scientific realism to fictional models is still on the table. In particular, Bokulich does not endorse mainstream realism;[2] rather, she opts for a 'moderate' kind of realism, i.e., one that is able to accommodate fictional model explanations alongside non-fictional ones. In my view, this modification is not required: instead, I propose to keep the mainstream notion of realism and modify the notion of representation. Then, my argument will be that fictional models may 'represent' in a partial sense aspects of a physical situation. I will base my argument on an understanding of the structural makeup of theories and models as an hierarchy of mathematical structures at different levels of abstraction. To justify this 'ladder of abstractions', as I call it, I will turn to James Clerk Maxwell's notions of 'physical analogies' and 'embodied mathematics', and Bokulich's own account of them in her (2014).

This paper is divided into five sections. In Section 2, I will offer a case study illustrating the role of a fictional model in the explanation of a quantum phenomenon. In Section 3, I present the view of Bokulich and the most relevant objections to it. Section 4 deals with the possible reassessment of her view. Conclusions are drawn in Section 5.

2. An Example: The Rydberg Atom

I begin by presenting a case where the explanatory role of a fictional model is manifest. It is the case of so-called Rydberg atoms, which is dealt with in detail by Alisa Bokulich (2008a). Rydberg atoms (named after Johannes Rydberg) are very simple quantum systems, consisting in light atoms that have been highly excited, so that their outermost electrons are at the threshold of ionisation. Their size becomes enormous, approaching the dimensions of minute macroscopic particles. Due to this fact, they are amenable to the methods of 'semiclassical physics'. In general terms, this means the employment of classical notions to study highly complex quantum systems, at the interface between the so-called microcosm and macrocosm. Faced with a lack of straightforward quantum mechanical solutions for such systems, scientists resort to hybrid models, seeking classical 'analogues' of the systems at hand and then mixing fictional features of a classical nature, mainly orbits traversed by imaginary particles, with genuinely quantum concepts such as wavefunctions and probability densities. Notable cases are those of 'quantum chaos', so named because the classical 'analogues' used in their study exhibit chaotic behaviour.

Rydberg atoms moreover offer fertile ground for philosophical considerations concerning the relations between classical and quantum. Bokulich compares such an atom with a grain of sand, remarking that

> These atoms call to mind Tom Stoppard's play Hapgood, in which he writes 'there is a straight ladder from the atom to the grain of sand, and the only real mystery in physics is the missing rung. Below it, [quantum] particle physics; above it, classical physics; but in between, metaphysics' […] As an atom that is the size of a grain of sand, Rydberg atoms are ideal tools for studying the 'metaphysics' of the relation between classical and quantum mechanics (Bokulich 2008a: 115).

[2] By 'mainstream realism' I mean the philosophical stance advocated by, e.g., Psillos 1999.

The historical precursor of the phenomena I am going to describe here is the so-called Zeeman effect, which concerns the changes in atomic spectra in the presence of a magnetic field. The effect consisted in the splitting of spectral lines into multiplets separated by spacings of variable size, with increasing complexity depending on the structural intricacies of the atoms and the strength of the magnetic field. In very simple cases, solutions were available, even based on the 'old' quantum mechanics. However, when the magnetic field used becomes sufficiently strong, complicated patterns appear that still defy a complete treatment by modern quantum mechanics. Notably, the dynamics of even the simplest atom, hydrogen, becomes classically chaotic when subjected to a very strong magnetic field (Bokulich 2008a: 115).

Rydberg atoms came into the picture with a number of experiments performed relatively recently, beginning in the late 1960s. Henceforth, my exposition follows Bokulich (2008a), to which I refer for details. In a series of experiments, researchers studied the spectra of barium atoms, reaching Rydberg states when excited through illumination with light. Increasing the intensity of the light, the spectra, as expected, showed peaks at the photon energies which could be absorbed by the atoms. When the photon energy exceeded the ionisation energy of the atoms, the peaks disappeared. A striking phenomenon occurred, however, when the experiments were repeated in the presence of very strong magnetic fields: the peaks persisted even after the ionisation energy was reached and passed.

A further complication was discovered, in similar experiments with hydrogen atoms at Bielefeld in the mid-80s that revealed irregular patterns of lines. Bokulich mentions the conclusion of the researchers involved, stressing the need to probe the connections of quantum mechanics with classical chaos utilising classical concepts (Bokulich 2008a: 117). Subsequently, the Bielefeld researchers achieved a breakthrough: they performed a Fourier transformation turning the energy dependence of the spectral patterns into a time dependence, with a striking result. A definite correspondence was revealed between the irregular spectral lines ('resonances') and hypothetical *classical* orbits of electrons in a *fictional classical model* of the same Rydberg atoms under the same conditions. Bokulich (2008a: 118) quotes the verdict of the scientists:

> In this work we have discovered the resonances to form a series of strikingly simple and regular organization, not previously anticipated or predicted […] The regular type resonances can be physically rationalized and explained by *classical* periodic orbits of the electron on closed trajectories starting at and returning to the proton as origin (Main et al. 1986: 2789–90; emphasis in original).

I want to stress two points in this discussion. So far, there seems to be an *explanation* of the Rydberg spectra in the above conditions based on the employment of a *fictional model*: the classical electron trajectories used do not exist. To repeat what I wrote in the Introduction, fictional models—in the sense in which I am using the term—are "theoretical structures describing physical systems that are not, in fact, instantiated" (Zorzato 2023). The classical model employed in the Rydberg atom case was discovered through analysis of experimental results, independently of quantum mechanics as the appropriate theory. However, justification was needed. This came with theoretical developments, resulting in the so-called

'closed orbit theory'. It established a kind of correspondence between "the average quantum density of states and the periods and stabilities of the classical periodic orbits, which allows a calculation of the quantum quantities on the basis of these classical quantities" (Bokulich 2008a: 120). In essence, it blended classical orbits with propagating waves that interfere and produce the observed patterns. There is definitely no question of quantum mechanics being reduced to, or replaced by classical mechanics. At the same time, the quantum mechanical wave interference considerations do not stand by themselves: the classical orbits are indispensable.

The second point I want to make is based on a surprising fact established in the late 1990s, when a group of researchers studied Rydberg atoms, of lithium in this case, in a strong electric field (Stark effect). Their result was that, starting from the experimentally observed spectrum, they managed to reconstruct the corresponding fictitious classical orbits. This sent the scientists wondering about the ontological status of the associated classical orbits, which undoubtedly were not real. It is this fact that underlies Bokulich's suggestion, that "What seems to be called for – given these experiments and the fertility of using classical trajectories in semiclassical mechanics more generally – is something less than a full-blown realism, yet more than a mere instrumentalism that dismisses them as nothing more than a calculational device" (Bokulich 2008a: 125).

3. The Tension: Fiction and Truth

We saw above how a fictional model involving classical electron trajectories plays an indispensable role in explaining a complex quantum phenomenon. This creates tensions for mainstream realism. Bokulich presents the problem in these terms:

> Science, it is commonly thought, must deal only in the truth, the whole truth (if possible), and nothing but the truth. After all, isn't fiction ultimately antithetical to truth? Won't scientists be misled into a labyrinth of confusion and be lulled by the mere illusion of understanding if they trade in fictions? Even those who have granted a limited function for fictions in science have denied that they can play a role in scientific explanation or in generating genuine knowledge. […] The difficulty, however, is that an examination of scientific practice reveals that models routinely play a central role in scientific explanation and that all models are non-veridical to some degree (Bokulich 2016: 2-3).

The issue that Bokulich addresses is a more general problem that arises in philosophy of science in dealing with models. Traditionally, there are two ways to interpret the extensive and fundamental role of models in science: the realist view and the instrumentalist view. According to the latter, scientific models are instruments, useful tools for predictions. Obviously, here the issue about the model's fictional nature does not arise. According to the former view, there is (some) correspondence of models with the world and with the entities they postulate, i.e. models genuinely represent their target system. This notion is essential for an explanatory role. One of the most influential notions of explanation is the one developed by C. Wesley Salomon (Salomon 1984a; 1984b; 1989), according to which an explanation is genuine if it describes the causal processes in the existing target system. This notion of causal explanation requires that the target system

exists in the world. In the case of fictional models, the realist's position is presented with the difficulty of how to account for the role of these models and their correlation with the real world.

Since fictional models are used in almost all fields of science, the tension with realism is spread over different contexts. Therefore, the importance of acknowledging the presence of this tension and of offering a solution is a requirement for philosophers and scientists alike. Let me now present Bokulich's argument in answering the challenge of fictional models. To reconcile the accepted fictionality of certain models with their recognised explanatory role in science, Bokulich offers an argument that accommodates the fictional nature of such models with a 'moderate' realism. Let us follow her argument step by step.

3.1. The 'Eikonic' Conception

In order to defend the explanatory feature of fictional models, Bokulich distinguishes the 'ontic' from the 'eikonic' conception of explanation. The ontic conception requires that "explanations are the concrete entities in the world" (Bokulich 2016: 1; 2018a). Even if not explicitly, the ontic conception shares its requirement with the causal notion of explanation developed by Salmon (Bokulich 2016: 5). Bokulich contrasts the ontic conception with what she calls the 'eikonic' conception. The eikonic approach is meant to allow non-causal explanations—i.e., fictional model explanations- alongside causal ones. It is based on three main points (Bokulich 2008a; 2008b; 2012; 2018a): first, the explanations must involve a scientific model. Second, the model doing the explanation has a counterfactual structure, in the sense that it is answering to 'what-if-things-had-been-different' questions. Third, not all fictional models explain: a 'justificatory step' is necessary to differentiate explanatory fictional models from non-explanatory ones. This step is understood as "specifying what the domain of applicability of the model is, and showing that the phenomenon in the real world to be explained falls within that domain".[3] It is a process which can either proceed from "an overarching theory, specifying the domain of applicability of the model", or instead "through various empirical investigations" (Bokulich 2011: 39). Therefore, the 'justificatory step' is an empirical question to be answered by scientists on a case-by-case basis. Since mainstream realism would not be in agreement with the eikonic approach, which allows for fictional model explanations, Bokulich proposes a slight modification of realism, to "moderate" it in some sense.

The main point is how to establish a structural correspondence between the model and the target system. According to Bokulich, "we require that the counterfactual structure of [the model] be isomorphic in the relevant respects to the counterfactual structure of [the phenomenon to be explained]" (Bokulich 2011: 39-43). As an example, Bokulich (2016) cites the explanation of the tides based on Newton's theory of gravity. Newtonian gravitation is considered a fiction in light of General Relativity. However, the Newtonian model—in virtue of the "similarity"[4] of the predictions of the Newtonian and the General Relativistic theories of gravity—is able to represent the tides, as well as the positions of the Sun,

[3] For more details, the reader is referred to the original papers of Bokulich (2011: 39).

[4] Alongside assertions that "General Relativity *exactly* reduces to Newtonian theory", it is stressed that "the Newtonian approach [...] is only valid (with justification from General Relativity)" under definite conditions (Mukhanov and Viatcheslav 2005: 10; 24; emphasis in original).

the Moon and the Earth along their orbits and along their possible variations (if, for instance, the Moon had a different mass, the model would explain the possible variation of the tides). The high precision of the model is, according to Bokulich, justified by the fact that it can describe the *explanandum* (the tides).

The structures of the model and the target are then isomorphic in the sense that they share in some way the same features. According to Bokulich, the real target has a structure and so does the explanation. The structure appears at different levels, both for the target and for the explanation of it. Appealing to Woodward and Hitchcock' account (2003: 198), Bokulich (2008a: 152) talks about the 'explanatory depth' of the model, i.e. "a measure of how much information the explanans provides about the system of interest" (Bokulich 2008a: 152). Bokulich, in detailed discussions of specific cases (e.g., Bokulich 2015), argues that a model can be associated with a relevant theory, in which case it stands in as a proxy for the theory when it captures generic features of the target system; it truly describes *aspects* of the target despite being fictional.

Two points should be stressed here. First, a fictional model may stand in as a proxy for a theory, but its role can be autonomous: the model does not ride piggyback on the theory. I'll return to this in the following. Secondly, the model "does aim to give genuine insight into the way the world is" (Bokulich 2011: 44); so it illuminates some genuine aspects of the target system which the relevant theory cannot. It can be that a theory may in principle explain the phenomena in a different way, even if in a more complicated way than the model itself. However, the model is necessary in cases where explanations based on the relevant theory are lacking. These points bring me to the criticisms that have been levelled at Bokulich. I shall argue that my own argument can counter both objections raised against Bokulich.

3.2. Criticisms

Samuel Schindler (2014) claims that Bokulich's aim at maintaining both the fictionality and the autonomy of the fictional models fails, unless she provides an extra argument for establishing the autonomy of fictional models. Commenting on quantum mechanical cases cited by Bokulich (2008a; 2008b; 2011; 2012), where the justification of fictional model explanations involves a 'link' with a relevant theory, Schindler writes:

> The tension is this: either model fictions are justified or they are not. If they are not, they provide no genuine explanation. […] But if the model fictions are justified, i.e., they are linked (in a very precise manner) to quantum mechanics through semi-classical theory […] how can model fictions be claimed to be explanatorily autonomous? (Schindler 2014).

The main point of Schindler's criticism is that the explanatory role is played by the theory and not by the fictional model. Thus, there is no reason to claim that the fictional model is explanatory because all the job is done by the theory and the model is merely a calculation tool.

James Nguyen (2021) too offers a criticism of Bokulich starting from one distinction: on the one hand, there are questions such as 'why does certain behaviour occur?'; on the other, questions like 'why does the counterfactual dependence

invoked to answer that question actually holds?'. According to the author, fictional models can answer the former but are in trouble with the latter (Nguyen 2021: 3229). But, if so, fictional models lose their fictionality, since the actual representation is the only one that remains. In Nguyen's words:

> [E]ither these models cannot answer these sorts of explanatory questions, precisely because they are fictional; or they can, but in a way that requires reinterpreting them such that they end up accurately representing the ontological basis of the counterfactual dependency, i.e., reinterpreting them so as to rob them of their fictional status. Thus, the existence of explanatory fictions does not put pressure on the idea that accurate representation of some aspect of a target system is a necessary condition on explaining that aspect (Nguyen 2021: 3229).

I will return to both criticisms in the following.

4. Reassessing Bokulich

Fictional models have a representational role with respect to a specific aspect of the associated theory's proper target. My argument for this claim hinges on what I dub 'the ladder of abstractions'. It can be captured by the slogan: the more you go up the ladder, the deeper you go into the object. What does it mean? The expression 'ladder of abstraction' is meant to highlight the hierarchical arrangement of mathematical structures making up a theory, or a model for that matter. To illustrate my point, I now turn to relevant aspects of J.C. Maxwell's work in developing his electromagnetic theory. At a certain stage in his endeavours, Maxwell made use of a mechanical model, which was fictional in my sense of the term:

> Maxwell constructed an imaginary physical system, contrived solely for the purpose of developing a mathematical scheme applicable to a specific physical domain. He could then draw consequences from this imaginary system to the physical domain of electromagnetism that was rich in experimental results (Hon et al. 2021: 253).

Bokulich's reading of J.C. Maxwell's method of using a mechanical fictional model points to his methodology of 'physical analogy' (Bokulich 2015): It is based on the use of an analogy to develop a new domain starting from a familiar one. The crucial point is that the analogy referred to is between the relations of things, not between the things themselves (Maxwell 1881: 52). On this basis, Maxwell developed his 'idle wheels' model (Maxwell [1861/62] 1890: 486). The core of the model was the use of a fluid, which was "not even a hypothetical fluid" but "merely a collection of imaginary properties" (Maxwell, 1890/1965: 160). Eventually, Maxwell reinterpreted the connexion between his mechanical contraption and his nascent electromagnetic theory, to demonstrate that the latter possessed generic features expressible in terms of the Lagrangian formalism of classical mechanics. Therefore, it could be embedded in that formalism in its abstract form (Maxwell [1876] 1890: 308).

Bokulich (2015) interprets Maxwell's methodology in terms of an hierarchical organisation of mathematically formulated theories addressing specific physical situations: at the highest level, there is the purely mathematical form (the

Lagrangian formalism). Below, there is a level of what Maxwell calls the 'embodiment' of that abstract mathematical form (Maxwell 1890/1965: 187). It is at this level that, according to Bokulich (2015: 31), a model can stand in as a proxy for a theory, in representing the target system. Generally, structures at various levels may be shared by different theories as well as models, even if those models are fictional. The hierarchy of mathematical structures is correlated to the abovementioned notion of 'explanatory depth' (Bokulich 201: 35).

I propose that the hierarchical structure indicates at which level *explanandum* and *explanans* are connected, and, depending on the level at which this happens, the explanation provided is more or less deep. Moreover, I suggest that fictional models can explain without even being directly related to a theory (such is the case of Bohr's atomic model, discussed in Bokulich 2008a). Indeed, it is possible for a fictional model not only to stand in as a proxy as claimed by Bokulich, but also to 'mediate' horizontally (Bokulich 2003) between different domains, establishing connections at higher levels of abstraction. Here, the role of physical analogies is evident. At the higher level of structural correspondences, mathematical structures are shared, allowing exploration and development of new domains. A model, even a fictional one, can capture essential features of a phenomenon targeted by an associated theory at a level below pure mathematics, that is, at the embodied mathematics level, where the model 'stands in as a proxy' for the theory.

In the process of probing the structure of the model, the depth of the explanation is also assessed. Indeed, the capacity of the abstraction is to be broader and to include more fundamental features, hence to reach deeper into the object, teasing out properties and relations of the target system. The less abstract the explanation, the more focused on the details of the phenomena it is. The success of fictional models is then explained by the range of abstraction achieved by the explanation: an adequate representation can succeed in providing physical insight into the target system, as the structure of the model is capturing *something* of the more abstract structural aspects of that system. To sum up, the 'ladder of abstractions' alludes to an hierarchy of mathematical structures as a fundamental feature of theory articulation. It is then possible to vary the degree of abstraction of the level of explanation, meaning that along the backbone of the ladder, the path of gaining knowledge depends on the level at which the explanation focuses on. Going upwards means going deeper into the object, zooming out to get the broader picture of its properties.

The 'ladder of abstraction' argument supports scientific realism because it allows capturing directly something of the object in the world. In this way, no modification of realism is required. Indeed, there is a correspondence between the *explanandum* and the *explanans* that satisfies the requirements of realism. In those cases when a fictional model is acting as a proxy for a theory, as in quantum mechanical situations, the representational role is inherited by the model because the structures shared with the theories represent an essential part of the theories' target. The crucial point here is precisely the possibility of high-level structures to be representational. In defence of this point, I claim that what fills the 'representational gap' for them is the physical interpretation that turns abstract mathematical relations into 'embodied' mathematics, which are in turn embedded into the full representation afforded by the theory.

As a result, through scrutinising in each case the concrete experimental and mathematical constraints that define the level at which structural correspondences between a fictional model and a theory obtain, scientists can tease out knowledge of physical connections inherent in the object of investigation but invisible to the proper theories concerned.

I turn next to the challenges posed by the criticisms of Bokulich's account. Schindler's criticism concerns a fictional model's autonomy in relation to a theory relevant to a concrete phenomenon. Autonomy is established in specific cases studied by Bokulich, where: (a) a fictional model can explain in the absence of any theory (Bohr' atomic model—Bokulich 2008a); (b) a fictional model is indispensably explaining features of quantum phenomena unaccounted for by quantum mechanics ('wavefunction scarring', quantum dots); and (c), in semi-classical physics, 'horizontal' models mediate between different sectors, constructed in manifestly autonomous ways (Bokulich 2003).

Concerning Nguyen's criticism, let me stress that, as I have already noted in relation to the Maxwell case, the 'target' of a *fictional* model is itself fictional, i.e., non-existent. However, the model acts as a proxy for a theory in virtue of encoding such properties of *that theory*'s actual *target system* as those entering in structural correspondences between the model and the theory. It is in this, and this sense only, that the model can be said to represent the theory's target, albeit in a restricted, partial way, although it is a *false* model of—i.e., *misrepresents*—that target in its totality (see Zorzato 2023).

5. Conclusion

Bokulich's contribution is a remarkable step towards the analysis and comprehension of the role of fictional models in science and in philosophy. Her philosophical approach shows that the instrumentalist position concerning the status of those models fails, since it has been proven that the explanations provided by fictional models are genuine. Her claim is justified on the ground of an isomorphism between the structure of the target and the structure of the model. The solution offered by Bokulich is a 'moderate' version of realism, that can accommodate both fictionality and the explanatory role of those models. However, according to the criticisms levelled at her approach, her argument does not show how the model can capture reality without being dependent on the theory, and it does not make clear how can a model be both explanatory and fictional. In my account, the approach of Bokulich can be reassessed in the spirit of mainstream realism. The main concepts I have considered are the notions of 'embedded mathematics' and of 'physical analogy' borrowed from Maxwell's works. Those two notions helped me articulate an analysis of the relation between the target system, the model and an associated theory that follows a process of moving along what I have labelled a 'ladder of abstraction'. Moreover, my approach helps dissipate the doubts arising from criticisms about the autonomy and the representational role of fictional models.

My proposal is well illustrated by the case of Rydberg atoms. Here, classical orbits are involved in the explanation of a quantum phenomenon. Following the 'ladder of abstraction' process, I claim that the structure of the classical orbits, even fictional, plays an explanatory role partially explaining the behaviour of the electrons. This is because a relation between the classical orbits and the density of quantum states is established in the context of 'closed orbit theory', amounting to

a structural correspondence at a definite level of abstraction, and equivalent to a certain depth in probing the phenomena. This argument answers the criticisms about the autonomy and the representational role of the model.

The main conclusion of my argument is that no modification of realism is needed. What I suggest is the need for a broader concept of representation, including representation of a system without representing that system in its totality. When the analysis is focused on the structure of the model at a higher level of abstraction, the ability of capturing some part of the structure of the target system is enhanced. The fictionality can be accommodated by the old, good scientific realism. The problem of the explanatory role of fictional models is a practical scientific issue and it is far from being covered yet. I hope that philosophers, on the basis of future scientific developments, will provide increasingly richer knowledge about the conditions for their use.[5]

References

Bokulich, A. 2003, "Horizontal Models: From Bakers to Cats", *Philosophy of Science*, 70, 3.

Bokulich, A. 2008a, *Reexamining the Quantum–Classical Relation: Beyond Reductionism and Pluralism*, Cambridge: Cambridge University Press.

Bokulich, A. 2008b, "Can Classical Structures Explain Quantum Phenomena?", *The British Journal for the Philosophy of Science*, 59, 217-35.

Bokulich, A. 2011, "How Scientific Models Can Explain", *Synthese*, 180, 33-45.

Bokulich, A. 2012, "Distinguishing Explanatory from Nonexplanatory Fictions", *Philosophy of Science*, 79, 5, 725-37.

Bokulich, A. 2015, "Maxwell, Helmholtz, and the Unreasonable Effectiveness of the Method of Physical Analogy", *Studies in History and Philosophy of Science*, Part A, 50, 28-37.

Bokulich, A. 2016, "Fiction as a Vehicle for Truth: Moving Beyond the Ontic Conception", *The Monist*, 99, 260-79.

Bokulich, A. 2018a, "Representing and Explaining: The Eikonic Conception of Scientific Explanation", *Philosophy of Science*, 85, 5, 793-805.

[5] This paper is based on work in a Reading Group at the History and Philosophy of Science Department of the University of Athens, conducted by Professor Stathis Psillos. I am indebted to all participants for insightful remarks and discussions. I particularly thank Vassilis Sakellariou for his support and advice. I am also grateful to the participants at the 3rd conference of P.R.I.N. in Urbino 2022 for their valuable suggestions.

My research was supported by the University of Tartu ASTRA Project PER ASPERA (European Regional Development Fund), by the grant of the Estonian Ministry of Education No IUT20-5 and the grants of Estonian Research Council No PRG 462 and PRG 492, and by the Personal Research Founding schemes under Grant IUT20-7, MOBERC14. The research for this article was conducted with the support from the European Regional Development Found (Dora Plus grant Action2).

The presentation at the Conference "Models, Structures and Representation" of June 20-21, 2022, University of Urbino, was supported by the Italian Ministry of Education, University and Research through the PRIN 2017 project "The Manifest Image and the Scientific Image" prot. 2017ZNWW7F_004.

Bokulich, A. 2020, "Losing Sight of the Forest for the Ψ: Beyond the Wavefunction Hegemony", in French, S. and Saatsi, J. (eds.), *Scientific Realism and the Quantum*, Oxford: Oxford University Press, 185-211.

Cat, J. 2001, "On Understanding: Maxwell on the Methods of Illustration and Scientific Metaphor", *Stud. Hist. Phil. Mod. Phys.*, 32, 3, 395-441.

Gelfert, A. 2017, "The Ontology of Models", in Magnani, L. and Bertolotti, T. (eds.), *Springer Handbook of Model-Based Science*, Dordrecht: Springer, 5-23.

Hendry, R. & Psillos, S. 2007, "How to Do Things with Theories: An Interactive View of Language and Models in science", in Brzeziński, J., Klawiter, A., Kuipers, T.A.F., Lastowski, K., Paprzycka, K., and Przybysz, P. (eds.), *The Courage of Doing Philosophy: Essays Dedicated to Leszek Nowak*, New York: Rodopi.

Hitchcock, C. and Woodward, J. 2003, "Explanatory Generalizations, Part II: Plumbing Explanatory Depth", *Noûs*, 37, 181- 99.

Hon, G. and Bernard, R.G. 2021, "Maxwell's Role in Turning the Concept of Model into the Methodology of Modeling", *Studies in History and Philosophy of Science,* 88, 321-33.

Kryukov, A. 2020, "Mathematics of the Classical and the Quantum", *J. Math. Phys.* 61, 082101.

Main, J., Weibusch, G., Holle, A., and Welge, K.H. 1986, "New quasi-Landau Structure of Highly Excited Atoms: The Hydrogen Atom", *Physical Review Letters*, 57: 2789-92.

Masoliver, J. and Ana, R. 2010, "From Classical to Quantum Mechanics through Optics", *European Journal of Physics,* 31,171-92.

Maxwell, J.C. 1881, *An Elementary Treatise on Electricity*, Oxford: Clarendon Press.

Maxwell, J.C. 1890/1965, "On Faraday's Lines of Force", in Niven, W. (ed.), *The Scientific Papers of James Clerk Maxwell*, I, New York: Dover, 155-229.

Meyer, H.D., O'Brien, C., Donald, P.F., Cox, K.C., and Kunz, D.P. 2021, "Optimal Atomic Quantum Sensing Using Electromagnetically-Induced-Transparency Readout", *Physical Review,* 104, 4, 043103.

Mukhanov, V. and Viatcheslav, M. 2005, *Physical Foundations of Cosmology*, New York: Cambridge University Press.

Nelson, E. 1966, "Derivation of the Schrodinger Equation from Newtonian Mechanics", *Physical Review,* 150, 4, 1079-85.

Nguyen, J. 2021, "Do Fictions Explain?", *Synthese*, 199, 3219-44.

Nölle, C. 2011, "Quantum Mechanics and Classical Trajectories", arXiv:1005.3786v2 [math-ph].

Psillos, S. 1999, *Scientific Realism: How Science Tracks Truth*, New York: Routledge.

Salis, F. 2019, "The New Fiction View of Models", *British Journal for the Philosophy of Science*, 72, 3.

Salmon C.W. 1984a, "Scientific Explanation: Three Basic Conceptions", *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association,* 293-305.

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

Salmon C.W. 1989, "Four Decades of Scientific Explanation", in Kitcher, P. and Salmon, W. (eds.), *Scientific Explanation*, *Minnesota Studies in the Philosophy of Science*, Vol. XIII, Minneapolis: University of Minnesota Press, 3-219.

Schindler, S. 2014, "Explanatory Fictions—For Real?", *Synthese*, 191, 1741-55.

Schrödinger, E. 1926, "Quantisierung als Eigenwertproblem", *Annalen der Physik,* 4, 79, 361-76.

Woodward, J. 2003, *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.

Zorzato, L. 2023 (forthcoming), "The Puzzle of Fictional Model", *Journal for General Philosophy of Science/Zeitschrift für Allgemeine Wissenschaftstheorie*, 1-12.

Empirical Success, Closeness to Evidence, and Approximation to the Truth

Gustavo Cevolani and Luca Tambolo***

* *IMT School for Advanced Studies Lucca*

** *Independent scholar*

Abstract

Realists and antirealists agree that different theories can be more or less empirically successful, even if they disagree on how to interpret this fact. Most of their arguments rely on how the notion of success is understood; still, few definitions of success are available, and their adequacy is doubtful. In this paper, we discuss some of these definitions and introduce a new measure of the success of a theory relative to a body of evidence aimed at overcoming some of their limitations. We moreover discuss how empirical success is connected to the approximate truth (or truthlikeness) of theories, a point of crucial importance for the defense of scientific realism.

Keywords: Empirical success, Evidence, Truthlikeness, Similarity, (Anti)realism.

1. Introduction

Today, the debate between scientific realists and antirealists is as lively and diverse as ever. A main point of contention is how to interpret the empirical success of our best theories: as a symptom of their approximate truth, as realists maintain, or instead as their ability to "save the phenomena", as antirealists suggest? One thing that both camps agree on, however, is the plain fact that theories can be, and often are, in fact, empirically successful, i.e., able to "account for" (fit, accommodate) a body of available evidence. It is moreover commonly assumed that in doing this, some theories may be better than others; in other words, that "empirical success" is a comparative notion, admitting of degrees.

In light of the above it is perhaps surprising that, as Malcolm Forster (2007: 589) notes, "[r]ealists and antirealists have not said much about how empirical success should be defined" (there are however important exceptions, discussed below). While much work has been devoted to defining adequate explications, e.g., of the empirical support or confirmation received by theories (Crupi 2021), or of their explanatory power relative to some evidence (Sprenger and Hartmann 2019: Ch. 7), it is not clear whether such notions are sufficient to exhaust that of success, and there are reasons to believe the contrary. In any case, defining success is not only interesting on its own, but also crucial to most of the arguments in the

Argumenta 9, 1 (2023): 73-88
ISSN 2465-2334
© 2023 Gustavo Cevolani and Luca Tambolo
DOI 10.14275/2465-2334/202317.cev

realism/antirealism debate. For instance, Laudan (1981) famously challenged scientific realists to explain in precise terms the link between the empirical success of a theory and its approximate truth or, to use his terminology, to justify the Downward Path (from approximate truth to empirical success) and the Upward Path (from empirical success to probable approximate truth). Of course, to accomplish such a task the realist needs an appropriate notion of success, one that can work together with the available explications of (probable) approximate truth (or truthlikeness, or verisimilitude). Antirealist equally cannot do without such a notion, at least if they want to be able to explain in what sense science develops and progresses toward increasingly successful theories (Niiniluoto 2019: Sect. 3).

This paper aims at systematizing some intuitions concerning various notions of empirical success found in the literature, in order to make explicit a couple of adequacy conditions that arguably should govern the use of the notion. In doing this, we hope to set some ground to further study, in a realist perspective, the links between success, on the one hand, and scientific progress as approximation to the truth, on the other. We start in Section 2 with a quick look at the current debate on realism and antirealism, emphasizing that both camps share the need for an adequate understanding of empirical success. In Section 3, we focus on three attempts to formally define such notion—due respectively to Hempel, Kuipers, and Zamora Bonilla—and point to some of their limitations. To overcome some of these, in Section 4 we introduce a new measure of success construed as the degree of similarity or closeness of a theory to the available evidence. Section 5 introduces the notion of the truthlikeness or verisimilitude of a theory and discusses how it interacts with empirical success as we propose to define it, in the light of Laudan's challenge. In Section 6 we offer some brief concluding remarks.

2. Why Success Matters

At the most general level, one can define scientific realism as "a positive epistemic attitude towards the content of our best theories and models, recommending belief in both observable and unobservable aspects of the world described by the sciences" (Chakravartty 2017). To be slightly more precise, realism is usually taken to be a package of views including (qualified versions of) three theses. The first is the metaphysical—or, depending on one's preferred parlance, ontological—thesis that the world that scientific theories aim to describe exists independently of our minds. The second is the semantic thesis that scientific theories, being attempts to describe the world and not just to systematize observations, make claims that must be taken literally, as having truth values. The third—the most interesting for our present purposes—is the epistemological thesis that our most successful sciences produce theories that offer approximately true descriptions of the aspects or fragments of the world that constitute their "targets".[1]

While the above package of theses is the backbone of any realist position, it seems fair to say that there are as many versions of scientific realism as there are scientific realists. It is not difficult to see why. Consider, for instance, an important issue raised by the realist's optimistic attitude towards our current best theories, from which embracing the epistemological thesis above follows naturally. Does such an attitude bring with it the commitment to the view that our best, most

[1] Such a characterization of realism follows quite closely the one proposed by Psillos (1999: XVIII) and is substantially equivalent to those suggested, among many others, by Niiniluoto (1999) and Chakravartty (2017).

successful theories get things basically right, and as a consequence, any mistake remaining in their descriptions of the world merely concerns matters of detail? Will the theories embraced by the scientists of the distant future be nothing but slightly amended versions of current most successful theories? According to Kyle Stanford (2015, 2021), a positive answer to such questions is what characterizes "classical" or "commonsense" realism, espoused in past decades by such authors as Smart (1968) and Putnam (1975).

Classical or commonsense realism, however, does not have much currency nowadays: in the majority of cases, Stanford points out, scientific realists today are more "historically sophisticated" than their predecessors, and therefore allow for the possibility of some future theoretical changes that will alter significantly the current scientific image of the world. As Stanford puts it, historically sophisticated realists take into proper account the revolutionary theoretical upheavals characterizing the past of science, and therefore tend to qualify their optimism by restricting it only to certain parts, or elements, of our current most successful theories. More specifically, so-called "selective realists" restrict their commitment to the parts or elements of our best theories that are responsible for their success, and that they maintain one can reliably identify (although different brands of selective realism differ concerning which parts of theories are deserving of realist commitment). Indeed, selective realism has become an important and lively tradition within the realist camp (see, e.g., Kitcher 1993, Psillos 1999, Cordero 2017, Alai 2021). But the qualified optimism of selective realists, as Stanford readily points out, is optimism nonetheless—that is, something that marks a difference between historically sophisticated realists, on the one hand, and antirealists, who do not embrace the epistemological thesis typical of realism, on the other hand.

One must mention, though, that full awareness of the track record of the scientific enterprise need not necessarily lead the realist to adopt a selective approach, or to be willing to concede that radical theory changes, analogous to the revolutionary ones that occurred in the past, will take place in the future. For instance, Fahrbach (2011, 2017, 2021) forcefully argues that our current best theories are of a different kind than past successful theories that have by now been discarded. In fact, our current best theories enjoy a much higher degree of success than theories of the past. On Fahrbach's account, this depends on both the quality and the quantity of the evidence supporting them, which is today enormously higher than it used to be in previous phases of the scientific development. In light of the extremely high level of support from the evidence enjoyed by our current best, most successful theories, Fahrbach suggests, the realist's embrace of the epistemological thesis is then by and large more justified today than it was in the past.

The above illustrates that the debate within the camp of scientific realism is today as lively and diverse as ever (see Alai 2017 and Saatsi 2018 for the state-of-the-art). More importantly for our present purposes, the preceding highlights that the notion of success—intuitively, the degree to which a theory or hypothesis H accounts for a body of evidence E—must play a key role within any viable realist position. Absent an appropriately defined notion of success, the epistemological thesis characterizing realism does not even make sense. In fact, the realist's optimism towards the content of our best theories and models hinges upon the fact that the realist views a theory's success as a (fallible) indicator of its (approximate) truth.

In "A Confutation of Convergent Realism" (1981: 32-36), Laudan famously challenged realists to show that there is in fact what he calls an "Upward Path",

namely, that a theory's success provides one with appropriate epistemic warrant for the theory's (approximate) truth. Many current versions of scientific realism have arguably been developed, at least in part, precisely in order to meet the challenge posed in Laudan's paper—a task that, of course, can only be accomplished with an adequate notion of success in hand. However, it would of course be a mistake to think that success matters only to realists.

Antirealists, a no less diverse crowd than that of realists, also need an adequate notion of success, in the absence of which it is impossible to make sense of the development of science. Think of the idea that there is scientific progress—that our current best theories are in some relevant sense better than previous, by now discarded, theories. Laudan (1977) has offered an antirealist characterization of scientific progress in terms of the increasing problem-solving effectiveness of theories. In order for such an account to work, a notion of success in problem-solving effectiveness is required. And even Kuhn, who viewed the development of science not as that of an enterprise getting nearer and nearer to "some goal set by nature in advance", but rather, "in terms of evolution from the community's state of knowledge at any given time" (1962/1970: 171), needed the notion of success to account for the theory-choices made by scientific communities.[2] To mention but one more instance of how antirealists too need an appropriate notion of success in order for their accounts of science to work, recall in what terms van Fraassen defines the aim of inquiry pursued by the constructive empiricist. Such an aim is empirical adequacy, where an empirically adequate theory is one that "saves the phenomena" in the sense that "what it says about the observable things and events in this world […] is true" (1980: 12). Of course, different theories may be more or less adequate in van Fraassen's sense, meaning that they will save a larger or smaller part of the phenomena; again, the notion of success in the sense in which we deal with it here is obviously involved.

In sum, the notion of success plays a central role in any viable account of science, be it realist or antirealist. Importantly, realists and antirealists need not disagree on the best way to characterize and measure success. To be sure, realists read into success something—as mentioned, a (fallible) indication of (approximate) truth—that antirealists maintain cannot be read into it. Still, both realists and antirealists agree, for instance, that success is a matter of degree: we need a comparative notion, since we want to be able to meaningfully say that a certain theory is more (less) successful than another (or as successful as another). Moreover, even the most optimist of realists readily agrees with antirealists that for most of the time scientists deal with theories that have already been falsified, and yet enjoy a certain degree of success, and must therefore be taken seriously nonetheless.

In the next two sections, we briefly review some selected attempts to provide rigorous explications of the notion of empirical success in the form of a measure of the degree of the success enjoyed by a hypothesis or theory with respect to some available body of evidence. Taking our cue from such attempts, we also present a new measure of success satisfying several adequacy conditions which arguably govern intuitive assessments of the relative success of different hypotheses.

[2] See Shan 2019 for a recent attempt to revive the accounts of progress put forward by Laudan and Kuhn.

3. Measuring Success

The effort of clarifying the links between success and approximate truth has led several scholars to develop formal accounts of both notions. These accounts provide rigorous definitions of the success enjoyed by a theory or hypothesis *H* relative to a body of evidence *E*, sometimes in the form of measures of such success (cf. Niiniluoto and Tuomela 1979: Ch. 7; Sprenger and Hartmann 2019: Ch. 7). Note that such accounts assume (in line with much discussion within general and formal philosophy of science) that it is possible to talk of the success of *H* with respect to *E* in a sufficiently general and abstract sense, i.e., not only relative to specific examples, scenarios or contexts of application.

A classic example of such an approach is Hempel's discussion of the notion of the "systematic power" of *H* with respect to a body of evidence or information *E*, first introduced in the last part of his celebrated 1948 paper on the logic of explanation (co-authored with Paul Oppenheim, reprinted in Hempel 1965). In Hempel's intentions, systematic power includes both the predictive and the explanatory performance of *H*, in agreement with the well-known thesis about the symmetry between explanation and prediction (Hempel 1965: 279). The intuition is that *H* has great systematic power when *H* "covers" a great part of the evidence *E*, in the sense that *H* entails a high proportion of the content of *E*. To make this precise, Hempel introduces an (epistemic) probability distribution p for the relevant language in which *H* and *E* are expressed (p is defined on the possible worlds that can be described by such language or, in Hempel's Carnapian jargon, on its constituents or state-descriptions). He then defines the *content* of a proposition *X* as the measure $cont(X) = 1 - p(X)$, in agreement with the intuition (shared by both Popper and Carnap, among others) that the greater the information content of *X*, the smaller its probability, i.e., the "size" of the set of possible worlds compatible with *X*. Finally, Hempel (1965: 287) defines the systematic power of *H* with respect to *E* as the ratio of the common content of *H* and *E* to the content of *E*:

1) $syst(H, E) = \frac{cont(H \vee E)}{cont(E)} = p(\neg H | \neg E)$

While Hempel introduces *syst*(*H*,*E*) as a measure of explanatory and predictive power, it is quite clear that it can be employed as a measure of the success of *H* on *E*; for instance, as noted by Ilkka Niiniluoto, *syst* can be used to formally explicate Laudan's notion of problem-solving effectiveness (Niiniluoto 1990: 438-39). Indeed, *syst* seems to capture well some intuitively sound conditions characterizing the notion of empirical success. As an example, if *H* deductively entails *E*, and in this sense it is maximally successful, then $p(H|\neg E) = 0$ and hence $syst(H, E) = 1 - p(H|\neg E)$ receives its maximum value 1, as expected. If *H* is tautological, it has no information content and, as such, it tells nothing about *E* since it entails no contingent consequences; accordingly, since $p(H|\neg E) = 1$, then *syst*(*H*,*E*) is minimal, i.e., 0.

An interesting consequence of Hempel's definition is that it allows us to compare falsified theories as far as their relative success is concerned. As we shall see in the following, this can be defended as an adequacy condition for any satisfactory explication of empirical success (Kuipers 2000: 94). If *H1* and *H2* are falsified by *E*, in the sense they both entail $\neg E$, they can still have different degrees of success. A more surprising consequence is that, as one can check, if *H1*, but not *H2*, is falsified by *E*, it could be that *syst*(*H1*,*E*) is greater than *syst*(*H2*,*E*). This is the

case, for instance, when *H1* is a highly informative but falsified theory, whereas *H2* is compatible with *E* but uninformative: in the extreme case where *H* is tautological, its systematic power is 0, i.e., the minimum.[3]

A less welcome consequence of Hempel's definition 1 above is however the following: if *H1* entails *H2*, then *H1* is always at least as successful as *H2*. Intuitively, this is because when *H1* is logically stronger than *H2*, it entails at least all the content of *H2* and perhaps more: accordingly, it cannot be less successful than *H2*. Formally, the reason is simply that if *H1* entails *H2*, then $p(H1|\neg E) < p(H2|\neg E)$ and hence $p(\neg H1|\neg E) > p(\neg H2|\neg E)$, which means that *syst* is always greater for *H1* than for *H2*. This result is troubling not only if empirical success is conceived as an indicator of (approximate) truth, but also if it is construed, more generally, as a cognitive utility guiding theory-choice, or it is used to define scientific progress, as Laudan suggests (cf. Niiniluoto 1990: 443). In fact, it implies that, if *H* enjoys some success relative to *E*, it is sufficient to add to *H* some piece of information *X* not already entailed by it to obtain a new theory *H* & *X* which is no less successful than *H*, even if the added information *X* is completely irrelevant or even false relative to *E.* In other words, increasing the empirical success of *H* becomes a "child's play": just strengthen *H* by conjoining it with any proposition *X* whatsoever (like "the Moon is made of green cheese").[4] In the extreme case, *X* can even be ¬*H*: in fact, success is maximized by an inconsistent, and hence maximally informative, theory.

The principle according to which success should co-vary with logical strength is highly problematic and, we argue, should be rejected as an adequacy condition for a measure of empirical success. However, it follows from Hempel's purely probabilistic account of the notion. Partially motivated by this problem, some have developed non-probabilistic explications of success. We shall briefly discuss two such accounts.

The first is due to Theo Kuipers who, in a series of works (Kuipers 1987, 2000, 2019), has defended a form of "constructive realism" based on a sophisticated analysis of the relationships between theories, evidence, and truth within a broadly "structuralist" framework. In doing so, he defines and discusses many different notions (like confirmation, progress, and truthlikeness) including that of empirical success. Cutting Kuipers' account to the bones, he distinguishes (following Laudan and others) between the "problems" and the "successes" of theories. Problems of *H* are established anomalies or counter-examples to *H*; successes of *H* are established facts that can be derived from it. Of course, *H* is the more successful, the more successes and the less problems it has. However, Kuipers is careful to emphasize that even if strictly speaking any counter-example to *H* falsifies it, this is not a sufficient reason to plainly reject *H* or consider it necessarily worse (in terms of empirical success) than a non-falsified theory. In his own

[3] In recent years, much work has been devoted to the logic of explanatory power, partly inspired by Hempel's early efforts (for the state-of-the-art, see Sprenger and Hartmann 2019: Ch. 7). Different probabilistic measures of the explanatory power of *H* with respect to *E* have been proposed, and interesting results about their axiomatic characterization and reciprocal relations obtained. Interestingly, none of these measures satisfies the requirement just discussed: if *E* falsifies *H*, the latter's degree of explanatory power is either undefined or minimal. This suggests that the notion of empirical success is richer than, even if connected to, that of explanatory power.

[4] This "child's play" objection is also standard in the truthlikeness literature, where it was raised against some earlier definitions of such notion (see, e.g., Kuipers 2000: 254).

words, theory evaluation has to "take falsified theories seriously" (Kuipers 2000: 94). This is reflected in his basic definition of comparative success (Kuipers 2000: 112, notation modified):

2) Theory *H1* is more successful than theory *H2* iff i) the set of problems of *H1* is a subset of that of *H2*; and ii) the set of successes of *H2* is a subset of that of *H1*; and iii) in at least one case the relevant subset is a proper subset.

In other words, if *H1* has at least one more success besides those of *H2*, or *H1* has at least one less problem than those of *H2*, *H1* is more successful than *H2*, and the shift from *H1* to *H2* counts as an instance of progress, understood as increasing success. As Kuipers notes, the assessment of the relative success of two (or more) different theories is always relative to a body of empirical evidence available at some point in time. Consequently, new evidence may always change the comparative judgment in the above definition.

It is worth noting that Kuipers' definition, just like Hempel's, allows one to compare falsified theories with respect to their relative success. If *H2* is falsified (i.e., its set of problems is not empty), *H1* may improve on it, for instance, by retaining all its problems and successes, and adding some more successes. In such a case, *H1* and *H2* are both falsified, but *H1* is more successful than *H2*. However, if *H1* is falsified and *H2* is not, *H1* cannot be more successful than *H2*, since in such a case, even if the set of successes of *H1* can properly include that of *H2*, the set of problems of *H1* cannot be a subset of that of *H2* (since the latter is empty and the former is not). Thus, Kuipers' basic definition does not satisfy the condition that falsified theories may be better than non-falsified ones, which is instead respected by Hempel's measure. On the other hand, if *H1* entails *H2*, then *H1* has all the problems and the success of *H2*; however, it could have no more successes and strictly more problems than *H1*, and so be less successful than it. Thus, Kuipers' definition satisfies the condition that empirical success does not necessarily co-vary with logical strength, a condition that Hempel's measure instead fails to meet. As we shall see in the next section, it is possible to define a notion of success very similar to Kuipers' one but eschewing the limits of both Kuipers' and Hempel's approaches.

Before turning to this, let us briefly discuss an account due to Jesús Zamora Bonilla (1992, 1996, 2000), providing another important step toward our own approach. Following Kuipers, Zamora Bonilla adopts a structuralist approach to theory representation. For our purposes, it is sufficient to focus on the simplest measure he discusses, which exhibits some interesting features. Zamora Bonilla introduces his measure as a measure of the estimated truthlikeness of a theory given the available evidence; as we suggest, it is more properly construed as a measure of "evidential similarity", i.e., as a measure of success defined as closeness to the empirically established truth (cf. Zamora Bonilla 1992: 347-49). The measure is defined as the product of the similarity $s(H,E)$ of theory H to evidence E and of the "rigor" $r(E)$ of the evidence, as follows (Zamora Bonilla 1996: 29; notation modified):

3) $evsim(H,E) \equiv s(H,E) \times r(E) \equiv \frac{p(H\&E)}{p(H \vee E)} \times \frac{1}{p(E)} = \frac{p(H|E)}{p(H \vee E)}$

Here, $s(H,E)$ measures, in probabilistic terms, the "overlap" between H and E; $r(E)$ is just the reciprocal of the probability of the evidence, taken as a measure of its informativeness. Measure *evsim* takes its maximum value if H entails E, i.e., when it is maximally successful, and has a number of other interesting features

(Zamora Bonilla 1996: 31ff). For our purposes, the main limitation of this account (that Zamora Bonilla carefully discusses in section 3 of his paper) is that it does not allow one to assess the relative success of falsified theories: in fact, if E falsifies H, then $evsim(H,E)$ is always 0, the minimal possible degree of success. On the other hand, we believe that Zamora Bonilla's account captures a crucial aspect of the notion of empirical success, i.e., that it should measure how "close" a theory is to the available evidence; his probabilistic measure $s(H,E)$, however, is too crude for this purpose. In the next section, we build upon this basic intuition in order to develop a more adequate notion of empirical success as similarity to evidence.

4. Success as Similarity to Evidence

The empirical success of H should depend on how well H accounts for the available evidence E. One natural way to spell out this intuition is defining the success of H on E in terms of the content of E which is also conveyed by H. As we saw in the previous section, Hempel's measure of systematic power does exactly this by employing a purely probabilistic notion of content (and hence of success), but it has some conceptual shortcomings. To avoid these, we suggest here another way of defining success, partially inspired by the proposals by Zamora Bonilla and Kuipers discussed above.[5]

The central idea is that H is the more successful the closer it is (in a suitably defined sense) to evidence E. To keep things simple, we rely on a quite minimal framework.[6] We assume that the evidence E is a collection of individual facts, each described by a single "basic proposition" of a finite propositional language. By "basic proposition" we mean an atomic proposition or its negation (in other words, basic propositions do not contain connectives except, possibly, for the negation). E can then be represented either as a set of m basic propositions or as their conjunction, the latter being the strongest evidential statement acceptable at a given moment in time. So, if A, B, C are atomic propositions, E could be expressed, for instance, both as $\{A, \neg B, C\}$ or as $A \,\&\, \neg B \,\&\, C$. Similarly, a theory or hypothesis H is a (consistent) collection or conjunction of k basic propositions of the same language. (Alternatively, one can think of such a collection as set of empirical consequences of a more complex theory at the observational level.)

The following terminology seems quite natural. Suppose that B is a basic proposition which appears as an element or conjunct of E. Then, we shall say that B is a (empirical) "match" of H if H entails B; that B is a (empirical) "mistake" of H if H entails the negation of B; and that B is a (empirical) "lacuna" of H if H does not entail B nor its negation. (Note that a lacuna, in this sense, is not an element or conjoint of H, but, so to speak, a "gap" of H with respect to E.) Intuitively, the matches of H count in favor of its empirical success; the mistakes and lacunae of

[5] A *caveat* may be relevant at this point. Following much of the literature, in this paper we leave on a side one important problem concerning success, i.e., the distinction between accommodation and prediction (which is crucial, e.g., in statistics, where success is defined as fit to the data). In other words, we are separating the problem of defining the success of a theory in terms of its matches (and mistakes) and that of defining when such matches are "genuine" or "fudged" (as with overfitting in statistics). The latter problem is carefully discussed by Forster (2007); for a very recent discussion of "predictivism", see Crupi 2023.

[6] The present framework is borrowed from the so-called basic feature approach to truthlikeness (Cevolani et al. 2011; Cevolani et al. 2013; Cevolani and Festa 2021) to be discussed in the next subsection.

H detract from it. More formally, let us denote with t_E (for "true with respect to *E*") and f_E (for "false with respect to *E*"), respectively, the number of empirical matches and mistakes of *H*. Then, we can define the following simple measure of the empirical success of *H* with respect to *E*:

4) $es(H,E) = \frac{t_E}{m} - \frac{f_E}{m}$

Recalling that *m* is the number of the evidential statements in *E*, *es*(*H*,*E*) amounts to the normalized difference between the number of matches and mistakes of *H*. Note that, even if the lacunae of *H* are not explicitly mentioned, they count against the empirical success by lowering *es*(*H*,*E*): if *H* has many lacunae, it cannot be much successful according to such measure. In the extreme case, when *H* entails no empirical consequence at all (i.e., it is an empty set or conjunction, with $k = 0$), it is completely "lacunose" (so to speak), and its degree of success is 0. In such case, with a slight abuse of language, we shall say that *H* is tautological, meaning that it entails no basic propositions at all.

As we argue, our simple definition satisfies several intuitive desiderata on the notion of empirical success. For instance, if *H* is "maximally successful" in the sense that it entails *E* (and hence *H* entails all the *m* conjuncts of *E*), then its degree of success is $es(H,E) = \frac{m}{m} = 1$, which is the maximum possible. On the other hand, if *H* has at least one mistake or one lacuna, then *es*(*H*,*E*) will be lower than 1. The minimal degree of success (i.e., –1) is reached when theory *H* entails the negations of all the *m* conjuncts of *E*, i.e., *H* is maximally unsuccessful; if *H* only makes mistakes, then its degree of success is always negative. In this connection, the degree of success of a tautology (in the sense defined above) is a sort of natural middle point: from a qualitative point of view, we could say that *H* is "successful" if $es(H,E) > 0$, "unsuccessful" if $es(H,E) < 0$, and "empirically neutral" otherwise. Note that, according to this simple measure, a non-tautological theory *H* counting exactly the same number of matches and mistakes has the same degree of success as a tautological one, i.e., 0.

It is also easy to check that our measure satisfies all the conditions discussed in the preceding section, thus allowing for simple assessments of relative success of different theories. In particular, it conveys as special cases Kuipers' comparative assessments of success: if *H1* has more matches and no more mistakes, or less mistakes and no more matches, than *H2*, then *H1* is more successful than *H2*. Moreover, *es* avoids the unwelcome consequence of Hempel's probabilistic measure. If *H1* entails *H2*, this does not imply that *H1* is more successful than *H2*. To see this, suppose that *H2* has only matches, and *H1* adds to these some mistakes: then *H1* entails *H2* but *H1* will be less successful than *H2*.

To sum up, we list below a number of conditions governing the notion of empirical success, which are satisfied by our measure. Without attempting here a detailed defense of all of these conditions, we suggest that they may work as adequacy conditions for any viable explication of success, or at least that they should be take into account when discussing one. Note that we do not claim originality concerning such conditions, partly borrowed from extant literature, and note also that we are not implying that they need to be logically independent or exhaustive. Assuming that *E* represents all the available evidence with respect to which two theories *H1* and *H2* are evaluated in terms of their relative success, we have:

ES1. If both *H1* and *H2* entail *E*, they are equally successful.

ES2. If *H1* entails *E*, and *H2* does not entail *E*, *H1* is more successful than *H2*.

ES3. If both *H1* and *H2* are falsified by *E*, they are not necessarily equally successful.
ES4. If *H1* is falsified by *E*, and *H2* is not falsified, *H1* may be more successful than *H2*.
ES5. If *H1* entails *H2*, *H1* may be more, equally, or less successful than *H2*.
ES6. If *E* entails both *H1* and *H2*, and *H1* entails *H2*, *H1* is at least as successful as *H2*.
ES7. If *H* entails *E*, and *E* entails *E'*, *H* is more successful on *E* than on *E'*.

Some comments are in order. The first two conditions deal with non-falsified theories, i.e., theories which are compatible with the evidence *E*. ES1 says, in a sense, that the best a theory can do is to fully entail the evidence: among such "maximally successful" theories, there is no difference as far as success is concerned.[7] ES2 says that maximally successful theories are more successful than theories that have mistakes or lacunae. The next two conditions concern instead theories that are falsified by the evidence. ES3 is the basic requirement that falsified theories are not all on the same level: it is possible to compare them according to their relative success. ES4 specifies that falsified theories may be even more successful than non-falsified ones (as discussed above in relation to Hempel's proposal). The next couple of conditions govern the relationships between success and logical strength. ES5 emphasizes that there is no general link: logically stronger theories may be more or less successful than weaker ones, depending on how they relate to *E*. However, in the rather special (and unrealistic) case where two theories are both verified by *E* (there are no mistakes, but only matches, for both *H1* and *H2*), the logically stronger is also the more successful. Finally, ES7 concerns the success of a single theory *H* with respect to two pieces of evidence: if *H* is fully successful on both of them, its degree of success will be higher on the more informative piece of evidence.

A full discussion and defense of ES1-ES7 will have to be left for another occasion. In what follows, we focus instead on some interesting methodological consequences of our definition of success. Before doing this, however, let us note a further, final point. A theory *H* can "go beyond the evidence" in the sense that it entails more (or different) empirical consequences than those that, collectively taken, form *E* (this happens for sure if *k* is greater than *m*). This implies that estimates of the success of *H* are always relative to the available body of evidence *E* and always revisable: if at a later time one discovers that some *B* (not already contained in *E*) is true (and hence becomes part of *E*), the empirical success of *H* relative to the new evidence may increase (if *B* is a match of *H*), decrease (if *B* is a mistake of *H*) or remain the same (if *B* is a lacuna of *H*).

5. From Success to (Expected) Truthlikeness, and Back

Our main reason for dealing with measures of empirical success like *es* is, as mentioned, to study the relationship between success, on the one hand, and truthlikeness, on the other hand, from a realist's point of view. Let us emphasize, however, that our *es* measure should be of interest also to the anti-realist. Indeed, anti-real-

[7] Of course, maximally successful theories may well differ under other, important respects, like their simplicity, unification power, etc.

ists need a way of comparing theories with respect to their relative empirical success, unless they are prepared to reduce all kinds of theory assessment to a binary, all-or-nothing classification of "successful" vs. "unsuccessful" theories (cf. Kuipers 2000: 94). Since all theories in the history of science (or at least those accepted or taken seriously for some time) probably had some degree of success, one can argue that the anti-realist needs at least a comparative notion of success obeying conditions ES1-ES7 above. Our measure *es* provides such a notion and, we suggest, is perfectly acceptable to anti-realists, since it does not involve any reference to truth beyond the evidence.

Having clarified this point, let us now turn to the idea of truth approximation. In a nutshell, a truthlike theory is one that provides much true information, and few false information, about its target domain. If a theory H is highly successful with respect to the available evidence, the realist feels confident that H is on the right track toward the truth. To put it differently, from the realist's point of view, the success of theories is a fallible, empirical indicator of their actual closeness to the (unknown) truth, and speaking of assessments of the relative truthlikeness of different theories is fully meaningful. To clarify these intuitions, however, the notion of truthlikeness needs to be defined in more details.

Interestingly, the same approach we adopted to define success can be used here to define (expected) truthlikeness (Cevolani et al. 2011; Cevolani et al. 2013; Cevolani and Festa 2021). Given a finite propositional language with n atomic propositions, the strongest true statement of such language will represent "the whole truth" about the target domain. (Of course, we assume that n is not smaller than either m and k.) This statement T is the conjunction of the n true basic propositions of the language. Intuitively, T is the most complete true description of the actual world, given the resources of our language; the other "constituents" of the language (conjunctions of n basic propositions) describe all the other possible worlds which are not actual (in total, there are 2^n constituents or possible worlds, including T). A theory H will be the more truthlike or verisimilar, the closer or more similar H is to T. In general, given a theory H and a constituent W, the similarity of H to W will be measured as:

$$5)\ sim(H,W) = \frac{t_W}{n} - \frac{f_W}{n}$$

i.e., as the normalized difference between the number of matches and mistakes of H with respect to W. Accordingly, the truthlikeness or verisimilitude of H is defined as:

$$6)\ vs(H) = sim(H,T) = \frac{t_T}{n} - \frac{f_T}{n} = \frac{t}{n} - \frac{f}{n}$$

where we can avoid the subscript "T" since here matches and mistakes are properly true and false, respectively. Note that the truthlikeness of H is maximal (and equal to 1) when H is the truth T itself; it is minimal (and equal to –1) when H is the conjunction of the negations of all basic truths. A tautology has 0 truthlikeness; a non-tautological theory is more or less verisimilar than it, depending on the balance of basic truths and falsehoods it entails: the more truths and the less falsehoods, the better in terms of closeness to the truth.[8]

[8] In this connection, one should note that we are employing here the simplest possible measure of truthlikeness proposed by Cevolani et al. 2011. In their more general account, the relative "weight" of truths and falsehoods in assessing the verisimilitude of H can be

Note that the truthlikeness *vs*(*H*) of *H* is well-defined only assuming that the whole truth *T* is actually known. Of course, this is not what happens in all interesting cases of scientific inquiry. Typically, an inquirer can at best rely on a body of evidence *E*, assumed to be true, and try to assess the estimated truthlikeness of different theories on the basis of such evidence. Such "educated guesses" about estimated truthlikeness are, for the realist, the best one can do by construing empirical success as a fallible indicator of the theory's "real" truthlikeness. To make this idea clear, we can follow Niiniluoto (1987, 2017) in defining estimated truthlikeness as the expected value of the actual truthlikeness of a theory. Assuming that a (epistemic) probability distribution *p* is defined on the possible worlds (constituents) of our language, we define the expected truthlikeness of *H* on *E* as follows:

7) $evs(H|E) = \sum_{W_i} sim(H, W_i) \times p(W_i|E)$

i.e., as the sum of the degrees of truthlikeness of *H* in each possible world W_i, weighted by its corresponding probability given the evidence *E*. In words, *evs*(*H*|*E*) is high when *H* is very close to (highly verisimilar in) the possible worlds that the evidence indicates as highly probable. Assuming that the evidence is veridical, *evs*(*H*|*E*) is a fallible estimation of *H*'s actual truthlikeness, that can be revised as new evidence becomes available. Note that, as evidence increases, such estimate becomes increasingly reliable; in the limit, when *E* singles out just one possible world (the actual one, described by *T*), the expected truthlikeness of *H* is the same as its actual truthlikeness.

Equipped with defensible explications of the notions of empirical success and (expected) truthlikeness—in the form of the measures *es*, *vs*, and *evs*—we can now deal with some issues of central importance in the debate between realists and antirealists. In particular, we can re-formulate Laudan's Downward and Upward Paths as follows (cf. Niiniluoto 1999: sections 6.4-6.5):

DP) If *H* is (highly) verisimilar, it is (highly) successful.

UP) If *H* is (highly) successful, it is expected to be (highly) verisimilar (its degree of expected truthlikeness is high).

These two principles provide a bi-directional link between the success of *H* and its (expected) truthlikeness (or probable approximate truth, in Laudan's jargon): if *H* is actually verisimilar (something we cannot ascertain), it should enjoy a high degree of empirical success; vice versa, if *H* is highly successful on *E*, then its degree of expected truthlikeness on *E* should be comparatively high. Laudan maintains that realists should accept in general both DP and UP, and should provide good arguments in their support. However, there are good reasons to think that these principles are too strong, and therefore realists need not commit to them (cf. Niiniluoto 1999, 2017, 2019). Indeed, one can show that both DP and UP are violated if the measures *es*, *vs*, and *evs* discussed in this paper are employed as adequate explications of the relevant notions.

In fact, one can prove that none of the following two principles (which are nothing but the 'translation' of Laudan's in our present framework) holds in general:

different, so that, for instance, the "loss" in verisimilitude due a mistake is greater than the "gain" due to a match. Here, for the sake of simplicity, we are instead assuming that matches and mistakes are equally weighted in assessing truthlikeness.

DP') If *vs*(*H*) is high, then *es*(*H*,*E*) is high.
UP') If *es*(*H*,*E*) is high, then *evs*(*H*|*E*) is high.

The main reason why these principles fail in general is very simple: as stated, they are completely silent on what *E* is, more specifically, on the quality of the evidence upon which the relevant assessments of success and expected truthlikeness are performed. As we saw in Section 4, however, the precise relationship between *H* and *E* (here construed as the closeness of *H* to *E*) is obviously crucial to assess the success of *H* on *E*. In other words, the information provided by *E* must play a crucial role in evaluating the links between success and (expected) truthlikeness—a role that DP' and UP' ignore altogether.

Two simple (if rather abstract) counterexamples will be sufficient to show why the two principles are untenable in general. Suppose first that *H* is highly verisimilar, meaning that *H* is very close to *T*, i.e., *H* has many matches and very few (or none) mistakes. (Of course, this is something that one cannot ascertain, and that we assume for the sake of the argument). Moreover, suppose that *E* is very uninformative, i.e., it entails very few evidential statements. It follows that the success of *H* on *E* could be very low, for the simple reason that *H* and *E* could well have very few elements in common, or even none if either *H* and *E* are "disjoint" or *E* is tautological. In other words, even if *vs*(*H*) is high (as assumed), *es*(*H*,*E*) can be very low (or even 0 in the extreme cases mentioned). This shows why DP' cannot hold in general. (To be sure, it can happen that *H* is highly verisimilar, *E* is uninformative in the sense just defined, and still *H* is highly successful on *E*, because it entails all the few elements of *E*; this, however, doesn't need to happen in general.)

As for UP', a similar counterargument can be given. Suppose, as before, that *E* is very uninformative, for instance because *E* consists just of one evidential statement *B.* Moreover, suppose that *H* not only entails such *B* (and hence it is maximally successful) but, as an extreme case, it is equivalent to *B* (and hence to *E*). In such case, *es*(*H*,*E*) is maximal, but *evs*(*H*|*E*) may be very low, especially if *n* is very high: in fact, *H* will be very uninformative, and hence cannot have a high degree of expected truthlikeness. In other words, UP' cannot hold in general.

The lesson to be drawn from the preceding discussion is that the evaluation of methodological principles like DP and UP cannot be made in general, but only on a case-by-case basis, by taking into account the specific body of evidence *E* available in the relevant context. This, however, does not mean that nothing can be said concerning the relations between success and truthlikeness. Indeed, a reformulation of principles DP and UP suggests itself as a possible way out of the counterexamples just discussed:

DP") If *E* is (highly) informative and *H* is (highly) verisimilar, then *H* is (highly) successful on *E*.
UP") If *E* is (highly) informative and *H* is (highly) successful on *E*, then *H* is expected to be (highly) verisimilar on *E* (its degree of expected truthlikeness on *E* is high).

These new conditions make clear the role of the evidence *E* presently available in assessing the link between the success of some theory *H* on that evidence and its (expected) truthlikeness. Of course, a rigorous formulation of DP" and UP" would require a formal explication of the "informativeness" of *E*, possibly in the form of some measure similar to the ones already discussed. This would allow

one to precisely formulate new principles—comparable to DP' and UP' above—and possibly to prove the existence of lower and upper bounds on the informativeness of E in relation to the success and (expected) truthlikeness of H. In this connection, we suggest that the framework presented here may be instrumental in proving that DP" and UP" actually hold under suitably defined conditions, a task that we have however to leave for the future.

6. Concluding Remarks

We started the paper by reviewing some aspects of the current debate between realists and antirealists, focusing in particular on the notion of empirical success of a theory or hypothesis H. That the relative success of different theories is a crucial ingredient of their evaluation and comparison is probably one of the few undisputed claims in such debate. In order to better assess the competing claims of realists and antirealists—and in particular the realist tenet that success is a fallible indicator of truthlikeness or approximate truth—we considered some formal explications of the notion of success, advanced by Hempel, Kuipers, and Zamora Bonilla. This led us to put forward a new definition of success (in the form of a measure defined on propositional languages) that, we argued, satisfies several adequacy conditions governing such notion, while overcoming the limitations of previous measures. In a nutshell, our definition construes the success of H as its similarity or closeness to the available body of evidence E.

In the final part of the paper, we showed how our account allows one to rigorously tackle some crucial aspects of the debate, and especially the discussion of the relationships between empirical success and (expected) truthlikeness. In this connection, our conclusions have been partly negative: one cannot prove, in general, strong "success theorems" (in the sense of Kuipers 1987, 2019) guaranteeing that high verisimilitude implies high success, or, vice versa, that high success implies high expected verisimilitude. In that sense, there is no general answer, on the part of the realist, to Laudan's challenge concerning the Upward and Downward Paths.

On a more positive note, we argued that Laudan's principles DP and UP are too strong, and therefore there is no reason for realists to embrace them without proper qualifications. This is because such principles ignore the issue of the quality of the available evidence, which becomes instead apparent in our account.

Moreover, such account has a number of advantages with respect to other proposals. First, it provides a defensible notion of success, satisfying a number of adequacy conditions discussed in the literature, but violated, at least in part, by other explications of success. Second, such notion is useful and perfectly acceptable also by the antirealist, thus providing a common ground for further discussion. Finally, our account provides a unified treatment of success (as closeness to observational truth) and of truthlikeness (as closeness to the whole truth) suggesting limited, but relevant, success theorems governing their relations. In this respect, further work is needed to explore both the potential and the limitations of the approach defended here.[9]

[9] We would like to thank Mario Alai, Enzo Crupi, Theo Kuipers, and two anonymous reviewers for precious comments on a previous draft. Gustavo Cevolani acknowledges financial support from the PRIN 2017 project ``From models to decisions'' (grant n. 201743F9YE) and the PRIN 2017 project "The Manifest Image and the Scientific Image" (grant n. 2017ZNWW7F_004).

References

Agazzi, E. 2017 (ed.), *Varieties of Scientific Realism: Objectivity and Truth in Science*, Cham: Springer.

Alai, M. 2017 "The Debates on Scientific Realism Today: Knowledge and Objectivity in Science", in Agazzi 2017: 19-47.

Alai, M. 2021, "The Historical Challenge to Realism and Essential Deployment", in Lyons, T.D. & Vickers, P. (eds.), *Contemporary Scientific Realism: The Challenge from the History of Science*, Oxford: Oxford University Press, 183-215.

Cevolani, G., Crupi, V., & Festa, R. 2011, "Verisimilitude and Belief Change for Conjunctive Theories", *Erkenntnis*, 75, 183-202.

Cevolani, G. & Festa, R. 2021, "Approaching Deterministic and Probabilistic Truth: A Unified Account", *Synthese*, 199, 11465-489.

Cevolani, G., Festa, R., & Kuipers, T.A.F. 2013, "Verisimilitude and Belief Change for Nomic Conjunctive Theories", *Synthese*, 190, 3307-24.

Chakravartty, A. 2017, "Scientific Realism", in Zalta, E. (ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2017 Edition.

Cordero, A. 2017, "Retention, Truth-Content and Selective Realism", in Agazzi 2017: 245-56.

Crupi, V. 2021, "Confirmation", in Zalta, E. (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2021 Edition.

Crupi, V. 2023, "The Case of Early Copernicanism: Epistemic Luck vs. Predictivist Vindication", preprint available at http://philsci-archive.pitt.edu/id/eprint/22583 (accessed 2023-11-03).

Fahrbach, L. 2011, "How the Growth of Science Ends Theory Change", *Synthese*, 180, 139-55.

Fahrbach, L. 2017, "Scientific Revolutions and the Explosion of Scientific Evidence", *Synthese*, 194, 5039-72.

Fahrbach, L. 2021, "We Think, They Thought: A Critique of the Pessimistic Meta-Meta-Induction", in Lyons, T.D. & Vickers, P. (eds.), *Contemporary Scientific Realism: The Challenge from the History of Science*, Oxford: Oxford University Press, 284-311.

Forster, M.R. 2007, "A Philosopher's Guide to Empirical Success", *Philosophy of Science*, 74, 588-600.

Hempel, C.G. and Oppenheim, P. 1948, "Studies in the Logic of Explanation", *Philosophy of Science*, 15, 2, 135-75.

Hempel, C. G. 1965, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Kitcher, P. 1993, *The Advancement of Science*, New York: Oxford University Press.

Kuhn, T.S. 1962/1970, *The Structure of Scientific Revolutions*, Chicago: The University of Chicago Press.

Kuipers, T.A.F. 1987, "A Structuralist Approach to Truthlikeness", in Kuipers, T.A.F. (ed.), *What Is Closer-to-the-Truth?*, Amsterdam: Rodopi, 79-99.

Kuipers, T.A.F. 2000. *From Instrumentalism to Constructive Realism*. Dordrecht: Springer.

Kuipers, T.A.F. 2019, *Nomic Truth Approximation Revisited*, New York: Springer.

Laudan, L. 1977, *Progress and Its Problems*, Berkeley and Los Angeles: The University of California Press.

Laudan, L. 1981, "A Confutation of Convergent Realism", *Philosophy of Science*, 48, 19-49.

Lyons, T.D. & Vickers, P. 2021 (eds.), *Contemporary Scientific Realism: The Challenge from the History of Science*, Oxford: Oxford University Press.

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

Niiniluoto, I. 1990, "Measuring the Success of Science", in *PSA 1990: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, East Lansing: The Philosophy of Science Association, 435-45.

Niiniluoto, I. 1999, *Critical Scientific Realism*, Oxford: Oxford University Press.

Niiniluoto, I. 2017, "Optimistic Realism about Scientific Progress", *Synthese*, 194, 3291-3309.

Niiniluoto, I. 2019, "Scientific Progress", in Zalta, E. (ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2019 Edition.

Psillos, S. 1999, *Scientific Realism: How Science Tracks Truth*, New York: Routledge.

Putnam, H. 1975, *Mathematics, Matter and Method*, Cambridge: Cambridge University Press.

Saatsi, J. 2018 (ed.), *Routledge Handbook of Scientific Realism*, New York: Routledge.

Shan, Y. 2019, "A New Functional Approach to Scientific Progress", *Philosophy of Science*, 86, 739-58.

Smart, J.J.C. 1968, *Between Science and Philosophy*, New York: Random House.

Sprenger, J. & Hartmann, S. 2019, *Bayesian Philosophy of Science*, New York: Oxford University Press.

Stanford, P.K. 2015, "Catastrophism, Uniformitarianism, and a Scientific Realism Debate that Makes a Difference", *Philosophy of Science*, 82, 867-78.

Stanford, P.K. 2021, "Realism, Instrumentalism, Particularism: A Middle Path Forward in the Scientific Realism Debate", in Lyons, T.D. & Vickers, P. (eds.), *Contemporary Scientific Realism: The Challenge from the History of Science*, Oxford: Oxford University Press, 216-38.

van Fraassen, B. 1980, *The Scientific Image*, Oxford: Oxford University Press.

Zamora Bonilla, J. 1992, "Truthlikeness without Truth: A Methodological Approach", *Synthese*, 93, 343-72.

Zamora Bonilla, J. 1996, "Verisimilitude, Structuralism, and Scientific Progress", *Erkenntnis*, 44, 25-47.

Zamora Bonilla, J. 2000, "Truthlikeness, Rationality and Scientific Method", *Synthese*, 122, 321-35.

The Representation of Reality in the Intelligent Use of Tools

Valentina Savojardo

University of Macerata

Abstract

Starting from some results of neuroscience, and especially of Embodied Cognition, I'll discuss the problem of the intelligent use of tools, as a useful perspective under which to investigate the link between common knowledge and scientific knowledge. The philosophical question from which I shall start my reflection is the following: how do we represent reality to ourselves when we intervene on it through the intelligent use of a tool? The answer to this problem will be developed in two fundamental steps. 1. The problem of the intelligent use of tools will be approached from the neuroscientific point of view of Embodied Cognition, from which, however, one risks drawing the impression of a radical separation between a common, practical knowledge and a more idealized scientific knowledge. 2. No such absolute separation exists, however, because all our representations of reality, when we intervene on it in a technical-practical sense, through the intelligent use of tools, depend on a collaboration between cognitive and motor elements of knowledge. This collaboration will be further exemplified through the Polanyian distinction between subsidiary and focal elements of knowledge, through which a functional mechanism can be identified, whereby knowledge is always mediated by action, both in our everyday activities and, at a more elaborated level, in science. Thus, a difference emerges, not in principle, but only in degree between common knowledge and scientific knowledge.

Keywords: Intelligent use of tools, Embodied cognition, Cognitive and motor elements of knowledge, Common knowledge and scientific knowledge.

> *With advancing age Renoir became crippled with arthritis. He lost the use both of his feet and hands; his fingers were immobilized in perpetual cramped rigidity. Yet Renoir went on painting for another twenty years until his death, with a brush fixed to his forearm. In this manner he produced a great number of pictures hardly distinguishable in quality or style from those he had painted before. The skill and the vision which he had developed and mastered by the use of his fingers, was no longer in his fingers.*
>
> (Polanyi 1958: 355, my italics)

1. Introduction

The quote I have chosen, taken from one of Michael Polanyi's major works, *Personal Knowledge*, seems to me to be particularly significant in introducing the

Argumenta 9, 1 (2023): 89-104
ISSN 2465-2334

© 2023 Valentina Savojardo
DOI 10.14275/2465-2334/202317.sav

themes my paper will focus on. The philosophical question from which my reflection starts is the following: how do we represent reality to ourselves when we intervene on it through the intelligent use of a tool?

The answer, in short, is to show how this representation occurs through a close collaboration between cognitive and sensorimotor elements of knowledge. Such collaboration emerges when we use a tool intelligently, both in our everyday 'practical' knowledge (common knowledge) and, at a more elaborate level, in scientific knowledge, which aims to produce controllable and sharable knowledge. When we speak of the intelligent use of tools, we mean all those situations, starting from our everyday actions up to the application of the most elaborate scientific and technological practices, in which we make use of tools that mediate the relationship between our body-mind system and the surrounding environment. If we consider, therefore, the relationship between cognitive and sensorimotor aspects of knowing in the intelligent use of tools, we come to deny, on the level of an epistemological critique, the difference in principle between a common, more technical, body-related knowledge and a more abstract, scientific knowledge. When we intervene on the reality, in a technical-practical sense, using particular tools, it is never possible to clearly separate theory from praxis (cf. in particular Buzzoni 1995, 2004, 2005 and 2008). In particular, this paper takes up and develops, extending them to the relationship between common knowledge and scientific knowledge, some considerations already presented in Buzzoni and Savojardo 2019.

Two fundamental steps, developed in the first and second paragraphs respectively, will be necessary in order to demonstrate the main thesis of this paper, according to which the nexus between motor and cognitive aspects in our representation of reality, when we intervene on it in a technical-practical sense, is an aspect that unites so-called common knowledge with scientific knowledge. Any radical in-principle separation between different types of knowledge, therefore, falls apart when we consider how we represent reality to ourselves when we intervene on it through the intelligent use of tools.

The first paragraph is intended to frame the problem of the relationship between motor and cognitive aspects of knowing from the perspective of Embodied Cognition, according to which cognitive activity depends not only on brain activity but also, and above all, on the action of the body on the mind (cf. in particular Rupert 2009, Shapiro 2010 and 2019). The nature of abilities in the intelligent use of tools is one of the most debated topics in this area: the solutions proposed at a scientific level are as diverse as the problematic nodes within the debate. As we shall see, the tendency of Embodied Cognition is to reduce the abilities related to the use of tools to the sensory-motor level (cf. Chao and Martin 2000, Grafton et al. 1997, Sakreida et al. 2016, Ferretti 2021, Iriki et al. 1996, Maravita and Iriki 2004), thus avoiding the opposition between motor and cognitive aspects, an opposition that nevertheless emerges in the face of some important challenges that Embodied Cognition cannot ignore. If, on the one hand, the use of familiar tools requires the retrieval of manipulative knowledge of a sense-motor nature, stored in our motor system, on the other hand, both the selection, creation and use of new tools, and the use of familiar tools employed in a new way, seems rather to require certain specific conceptual skills (Caruana and Cuccio 2015). The use of certain purely cognitive functions of a causal or inferential nature—referred to as 'technical reasoning' (Osiurak et al. 2010) or 'mechanical problem solving' (Goldenberg and Hagman 1998)—seems necessary.

From this debate, an important distinction emerges between a knowledge that we could define as sensorimotor, mostly linked to the use of familiar tools, in certain particular situations in which the tool ends up implying a change in the sensory system in which it is incorporated, and a more abstract knowledge that concerns the objective relations that apply between the objects themselves, regardless of our particular sense organs and the context of interests and meanings in which the objects are used (cf. in particular Osiurak 2014 and Goldenberg 2013). In the face of such neuroscientific findings, what can philosophical reflection say?[1] Developing in the light of the empirical data provided by scientists, on a philosophical-epistemological level, the risk emerges that the tension between hypotheses about cognitive and sensorimotor abilities in the intelligent use of tools, could turn into a form of dualism between distinct types of knowledge. On the one hand, there would be a common knowledge, which we can find in the use of familiar tools, a 'practical' knowledge that accompanies us, mostly unconsciously and automatically, in our daily activities, and on the other hand, a more abstract scientific knowledge, which concerns the objective relations between physical objects and which seems to be mostly about the invention of new tools or the use of familiar tools in an original way.

The second part of this paper will show that this distinction in principle cannot apply in our technical-practical intervention in reality, in which it is not possible to separate thought from action, because in it the use of any tool always becomes an intelligent, conceptually mediated use. Technical and practical elements linked to the use of our body intertwine with cognitive elements, as we try to focus on an aspect of reality, intervening on it through a tool. This applies in the context of common knowledge, as in science. The principled distinction between the two fields, therefore, no longer makes sense. In order to support this argument, I will finally refer to the Polanyian distinction between subsidiary and focal elements of knowledge. This distinction is in fact taken up by Polanyi himself in order to clarify the use of tools that we commonly assimilate to our body in order to carry out certain technical-practical operations, both in our everyday life and, at a more elaborate level, in scientific practice. Without going into M. Polanyi's thought in depth, reference to his epistemology of the human person will be useful to clarify the link, of unity, on the one hand, and distinction, on the other, between the motor and cognitive aspects of knowledge. On the one hand, indeed, with respect to our technical intervention in reality, it is necessary to deny the clear separation, in an ontological sense, between two spheres of knowing, one practical, linked to the body, and one more abstract, linked to the action of the mind; on the other hand, however, this necessity does not prevent us from distinguishing, in a sense that can be said to be functional, two perspectives on reality.

Bearing in mind the Polanyian proposal, we speak of a functional distinction with reference to the knowing subject who, in the performance of any practical activity, as in the use of particular tools, in our everyday life, or in science, can choose whether to direct his or her focal, immediate attention to the so-called

[1] By examining the representation of reality in the intelligent use of tools, this paper is part of the collaboration between philosophy and cognitive sciences (cf. expecially Bennett et al. 2007 and Bennett and Hacker 2022). If indeed, on the one hand, cognitive sciences open up the study of certain mental processes to empirical investigation, on the other hand, philosophy has the task of questioning the tools for investigating these processes, highlighting their limits and potential.

subsidiary elements, mostly linked to the use and involvement of the body; or the subject can act from these elements, incorporating them or integrating them in an almost automatic way into his or her complex body-mind system.

The problem of the intelligent use of tools, investigated in the context of Embodied Cognition, may thus be considered a paradigmatic case useful in showing the link between motor and cognitive aspects of knowledge, and thus the link, in a more general sense, between common knowledge and scientific knowledge.[2]

2. Embodied Cognition and the Intelligent Use of the Tool[3]

Our ability to use everyday tools requires different skills and is today a topic of great interest not only for cognitive psychology but also for philosophy and neuroscience. In particular, the problem of the nature and the role of the abilities involved in the intelligent use of tools represent a challenge for Embodied Cognition,[4] which aims at investigating the mutual dependence between body and mind, re-evaluating, compared to traditional cognitive theories, the role of the body in the different cognitive functions. Embodied Cognition is distinct from (but also closely intertwined with) three other research paradigms: those according to which the mind must be considered not only as 'embodied', but also as 'embedded', in both a natural and cultural context (Hutchins 1995), 'extended', i.e. extended to its instrumental extensions (Clark and Chalmers 1998, Wilson 2004, Menary 2010), and 'enactive', i.e. capable, through its action, of perceiving and structuring the world in which it finds itself (Varela, Thompson and Rosch 1991, Noë 2004 and Thompson 2007). From these perspectives, the cognitive system is not about a disenchanted Cartesian mind that manipulates symbols, it is based on human interaction with the physical, cultural and social dimensions of the world.

In Embodied Cognition, the answers to the problem of intelligent tool use have been concentrated around two opposite poles. The prevailing tendency has been to attribute skills related to the use of tools to the sensorimotor level, putting more cognitive skills in the background. On the contrary, a second trend, especially to explain the new and original use of tools, has considered it necessary to introduce types of reasoning that would be based on the acquisition of abstract mechanical laws, at least partially independent from the functioning of the motor system. The use of tools seems to represent a capacity situated halfway between sensorimotor skills and more abstract cognitive skills.

The first trend can be seen, for example, in two of the main answers that neuroscientists, in Embodied Cognition, have provided to the problem of the

[2] It may perhaps be useful to note that this article neither intends to distinguish between sensorimotor knowledge on the one hand and a more abstract knowledge of the physical characteristics of objects on the other (see for this distinction Osiurak 2014 and Goldenberg 2013), nor to identify these two types of knowledge with, respectively, common knowledge and scientific knowledge. What is at stake here is only to highlight how some problems that have emerged from the debate within neuroscience may cause philosophical reflection to run the risk of a dualism between two types of knowledge—one common, more practical, and one scientific, more abstract—dualism that is not defensible if we think about the way we represent reality by intervening in it through the intelligent use of tools.

[3] On this point, see also Buzzoni and Savojardo 2019.

[4] For a general overview of the topic, see the following texts: Shapiro 2019 and Palmiero and Borsellino 2018.

intelligent use of tools: that of *affordances* and that based on the concept of *embodiment* of the tool in the subject's motor schema.

According to the theory of affordances (see e.g. Chao and Martin 2000, Grafton et al. 1997, Sakreida et al. 2016, Ferretti 2021), initially inspired by Gibson (1979), the observation of the characteristics of a tool is able to evoke the motor programme necessary for its use. The characteristics of an object suggest to the agent the appropriate way to use the observed object: the affordances theory is therefore based on the necessary agent/object relationship. As has been observed, however, this relationship implies a reference to further, equally necessary relationships between the instrument and the structural characteristics of the objects and materials with which the instrument relates. Already with reference to the theory of affordances, certain problematic aspects have been stressed in the literature. If it is true, in fact, that the affordances of an object determine the agent/object relationship by enacting a series of transformations at the visual-motor level, the other relationships involved in the intelligent use of tools, such as the relationship between the object and other objects, or the different ways in which a tool can be used, cannot be explained through the affordances theory alone, as such operations seem to require further work at the semantic-cognitive level (cf. Caruana and Cuccio 2015).

A second sensorimotor theory supported in the field of Embodied Cognition is the one founded on the embodiment of the instrument in the subject's motor schema (see e.g. Iriki et al. 1996). This theory is based on the idea that the use of an instrument implies a change in the sensory and motor system in which the instrument itself is embedded. The tool thus becomes part of a new physical entity; hence the idea that the use of tools requires, rather than a complex series of cognitive elaborations, a plastic body schema, capable of incorporating external elements into itself (for a review, see first of all Maravita and Iriki 2004, but see also the following works: Berlucchi and Aglioti 2010, Johnson-Frey 2003, Cardinali et al. 2009, Caruana 2012).

However, the insistence on the sensorimotor aspect with which these theories have often been supported has provoked, in reaction, an opposite trend. Studies that have provided results in favour of the existence of an affordance effect, for instance, have shown that the latter is nevertheless conditioned by perceptual selection processes (cf. Makris, Hadar and Kielan 2013). Multiple experiments, moreover, have shown that what an individual intends to do with an object, i.e., the goal he or she has in mind, changes the hand attitudes during the movement to grasp the object (see e.g., Sartori, Straulino and Castiello 2011, Caruana and Cuccio 2015). Above all, while sensorimotor skills prevail in the case of standard use of familiar tools, in the cases of using new tools on the basis of analogy with known procedures and in the case of using known tools according to new procedures, mental operations seem to be involved which, although connected to the motor system, cannot be traced back to it without residue. Familiar tool use naturally always requires, at least to some extent, a set of sensorimotor skills, but the finding that certain brain damage is more significantly correlated with difficulties in both using new tools and using old tools in a new way, rather than with using familiar tools, has been deemed sufficient to postulate the existence of particular cognitive skills (Goldenberg and Spatt 2009: 1653).

There is a common tendency to consider the ability to use certain objects in an original way as if they were particular tools (a coin as a screwdriver) or the ability to use certain tools in an unconventional way (a fork as a comb) as

evidence of intelligence. These are in fact actions that require a certain amount of reasoning about the structural and mechanical characteristics of the object:

> A basic requisite for detecting non-prototypical uses of common tools or possible uses of novel tools is recognition of structural properties which determine the possibilities and limits of mechanical interaction with other objects. For using a coin to replace a screwdriver, flatness and rigidity are decisive structural properties. Flatness permits insertion of the coin into the slot of the screw, and rigidity secures transmission of rotation from the hand via the coin to the screw (Goldenberg and Spatt 2009: 1646).

In particular, knowledge of the structural properties of a tool is primarily concerned with the interactions of the tool with other objects or materials, rather than with the relationship between the object and the acting subject. For these reasons, the concepts of 'mechanical problem solving' (Goldenberg and Hagman, 1998), 'mechanical reasoning' (Hegarty 2004) and 'technical reasoning' (Osiurak et al. 2009, Osiurak et al. 2010, Osiurak 2014) have been introduced. They would all be based on the acquisition of abstract mechanical laws, at least partially independent of the functioning of the motor system, and would easily explain the paradigmatic case of the unconventional and new use of already known tools.

Now, a careful examination of some of the pages or assertions of the participants in this debate shows that, although we are here predominantly faced with a tension between empirical hypotheses that tend to be opposed with respect to the solution of a particular scientific problem, there is in some cases, at a properly philosophical-epistemological level, the unconscious introduction of a certain naturalistic reductionism or philosophical dualism, respectively. The distinction between sensorimotor knowledge and a more abstract knowledge of the general principles of physics and mechanics can be illustrated by two examples taken from two different authors.

According to Osiurak "sensorimotor knowledge is supposed to contain information about the usual manipulation of tools (egocentric, user-tool relationship), and not about the objects with which they are usually used (allocentric, toll-object relationship)" (Osiurak 2014: 91). In other words, on the one hand, there is knowledge that is directly dependent on and related to our interests and the concrete and particular situations in which we find ourselves, and on the other hand there is knowledge that concerns objects as such, and thus abstract and universal knowledge, or knowledge that is valid in itself; on the one hand, knowledge that has to do with the particular as the direct object of our cognitive and practical interest, and on the other hand, knowledge that examines the objective relations between the objects themselves, regardless of our particular organs of sense and the context of interests and meanings in which we use them.

This opposition (which can easily be related to the old dichotomy between things for us and things in themselves) can also be found, albeit more indirectly, in Goldenberg. He introduces a so-called intermediate knowledge that accompanies us throughout our lives and is acquired in and through our moving in a three-dimensional world occupied by solid objects (cf. Goldenberg 2013).

Now, the introduction of an intermediate term in no way attenuates the epistemological opposition presupposed here between a sensorimotor knowledge in particular or individual situations, which properly concerns the use of familiar tools, and a knowledge that, to use a passage quoted by Goldenberg and Spatt

2009, contains "the comprehension of mechanical interactions of the tool with other tools, recipients or material" (Goldenberg and Spatt 2009: 1653), that is, an abstract and idealised knowledge.

Despite the fact that these authors speak of a cooperation between the two types of knowledge, the distinction between the two fields is repeatedly stressed and risks appearing as a qualitative and principled difference between different forms of knowledge.

If the dualism between a sensorimotor knowledge that seems to be bound to the body, and a more abstract knowledge aimed at mechanical laws that concern objects considered in themselves, loses all contextual relativity, we end up presupposing, at a philosophical-epistemological level, a distinction between two cognitive domains that is in no way tenable for the epistemological reasons we will examine shortly.

Neuroscientific findings around the problem of the intelligent use of tools, in the specific field of Embodied Cognition, constitute the starting point for philosophical reflection. As we shall see, in fact, at the philosophical-epistemological level, a clear separation between the following two types of knowledge is not tenable: a practical or technical knowledge that accompanies us in an almost automatic or unconscious manner in our daily activities, and a scientific knowledge, which specifically concerns the objective characteristics of the objects we use and aims at complete intersubjective controllability.

3. The Representation of Reality and our Practical-Technical Intervention in It

Starting from the debate on the intelligent use of tools in Embodied Cognition, in the light of some important experimental results (cf. especially. Brandi et al. 2014, Valyeaar et al. 2007, Osiurak et al. 2010, Goldenberg and Spatt 2009), it is evident how difficult it is for neuroscientists to succeed in defining the relationship between cognitive functions and the sensorimotor functions that determine the intelligent use of tools by human beings. As I have said, the risk, at the philosophical-epistemological level, is that of arriving at a principled difference between a technical or practical knowledge linked to the body and a more abstract and objective scientific knowledge.

The purpose of this paragraph is to show that this difference is not tenable if we consider our technical-practical intervention in reality an intervention that often makes use of particular tools, both in our everyday activities and in science. As we shall see, however, this statement does not prevent us from understanding the distinction between these spheres in a new sense, not as a clear separation of principle, but as a difference of perspectives on the same reality.

When we make use of any instrument, from the stick to move in the dark to the probe to explore space, we do so with the aim of intervening in the reality around us, guided by an underlying intention that may be that of seeking the exit from the dark room we find ourselves in or that of getting to know new aspects of the spatial universe. In this sense, the use of an instrument that mediates between our body and reality is always an intelligent use and this presupposes an important link between thought and action, and between cognitive and motor elements of knowledge, in our technical-operational intervention on reality.

To make this point clearer, let us start with experimental science. The general idea is that the theoretical moment and the technical moment are two aspects that

can be distinguished in experimental science only on the level of reflection, because, on the one hand, in the concreteness of doing science, the theoretical moment is the condition of possibility of the knowledge of certain aspects of reality and of possible causal links that can be resolved, in principle, in technical applications accessible to the entire scientific community; on the other hand, the technical moment possesses truthful relevance when it translates into conceptually mediated actions (cf. Buzzoni 2008: 24-25). There is no human knowledge that is absolutely non-technical, just as there can be no knowledge that is merely practical-technical, unmediated by concept. This means that any attempt to epistemologically separate pure, abstract or idealised science from its practical applications is doomed to failure. Knowledge of empirical reality cannot be separated from a practical or instrumental intervention in nature, an intervention that, in turn, is always mediated by the concept, without which action could not be distinguished from mere chance occurrence.

In support of this argument, we consider the role of counterfactual assumptions, which outline a series of conditionals present in science (see especially Williamson 2016 and 2020), as in common thought. In everyday life, in fact, the mind often constructs possible alternative scenarios to real situations, scenarios that allow the agent to move in the real world, for example, as some empirical research has also shown, through a type of reasoning by opposites. According to some recent studies in cognitive psychology (see in particular Branchini et al. 2016, 2021, Bianchi and Savardi 2006, Bianchi et al. 2017a, b, 2020, Byrne 2016, 2018, Dumas et al. 2013, Evans 2007), the role of opposites should in fact be understood as a general organising principle of the human mind. Interestingly, it is also able to represent a certain perceptual datum by hypothetically excluding other possibilities, which are not directly perceived by the senses: it is possible, for example, to perceive the red of a rose, the object of direct observation, hypothetically assuming the possibility that it could be another colour, and then rejecting this possibility on the basis of the relationship between my eyes and the object. Now, without this hypothetical capacity of the mind, our techniques of intervention in the reality would be indistinguishable from the simple natural change of things. Our reasoning in a counterfactual manner becomes the condition of our interventions on the real, showing different cause-effect links in empirical reality from time to time. Certainly the same mental processes that we use in our daily lives also apply to scientific thinking, albeit at a more elaborate cognitive level: without the construction of counterfactual scenarios, the scientist could not intervene in reality in any way. Like the historian, the natural scientist too, in order to explain a certain event, must ask oneself what might have happened in hypothetically different situations (for such considerations see especially Buzzoni 2008: 116-117).

When we use any tool, cognitive and motor elements work together, in the development of a knowledge that is also always acting. But if on the one hand we cannot accept the difference in principle between two separate cognitive spheres because, as we represent reality in our technical intervention in it, our thinking necessarily translates into shared practices; on the other hand, the distinction between cognitive and motor elements of knowing in the use of tools can be reconsidered by examining the distinction between subsidiary and focal elements of knowing by the Hungarian philosopher M. Polanyi. The relationship between these elements is, in fact, used by Polanyi both to exemplify the mechanisms underlying the intelligent use of tools and to clarify the body-mind relationship.

In order to understand the meaning of the distinction between subsidiary and focal elements of knowing, and how this distinction can be useful in clarifying the link between cognitive and motor aspects of knowing always mediated by action, it is necessary to introduce the Polanyian concept of 'tacit knowledge'.

Even scientific knowledge, which seems at first sight to present itself as completely explicit knowledge, according to Polanyi, contains a 'tacit' or 'unexpressed' moment, connected to pre- or a-linguistic skills. The role assumed by such skills in the scientific enterprise is, however, a serious problem (cf. Buzzoni and Savojardo 2019 and Savojardo 2013). If we understand these abilities as something in principle inexpressible in the form of verbal and discursive knowledge, they end up being part of an obscure background inaccessible to rational reconstruction. One would arrive, in this sense, at an ontological distinction between two realities, one expressible and the other unexpressed, tacit, or in any case not completely translatable on a conceptual level. In this sense, one cannot accept, from an epistemological point of view, the presence of a logical or explanatory vacuum in scientific knowledge, which by definition must be an intersubjectively controllable and reconstructible knowledge in every step. If, on the other hand, the distinction between tacit and explicit is understood in a functional sense, as if the transition from one sphere to the other coincided with a change of perspective on the actual data, then it is possible to think of science as always being connected to implicit knowledge that can in any case, in principle, become explicit.[5] Thus not only can an implicit ability be made explicit, but also an explicit ability can become implicit and operate at an unconscious level, in a circular but always renewed relationship between tacit and explicit.

In order to clarify how the relationship between tacit and articulate knowledge can be understood in a functional sense, we can turn to the studies of Gestalt psychology on perception, following the Polayian proposal and the distinction between subsidiary and focal awareness of the details of an object.

Polanyi identifies a 'logic of tacit inference' in the example of perception and the figure-background relationship through which we are able to focus on an object in front of us: "Every time we concentrate our attention on the particulars of a comprehensive entity, our sense of its coherent existence is temporarily weakened; and every time we move in the opposite direction towards a fuller awareness of the whole, the particulars tend to become submerged in the whole" (Polanyi 1969: 125).

Now, what is true for the attention paid to details, which risks making us lose the meaning of the whole, is also true for the abilities connected to the use of our body, which tend to become paralysed if the gaze of the person performing them is directed at single bodily movements: a pianist who shifts his attention to his fingers while playing risks becoming confused and will be forced to interrupt his performance. However, it is thanks to the details, seen as a whole, that we are

[5] There is an important oscillation in Polanyian thought with respect to the role of tacit ability. On the one hand, in fact, perhaps also due to the polemical intent towards logical empiricism, Polanyi sometimes seems to affirm that 'abilities' are in principle inexpressible in the form of verbal-discursive knowledge. On the other hand, Polanyi does not understand the distinction between tacit and conscious abilities as an ontological distinction, but rather as a distinction of a properly functional kind. In this case, the distinction between tacit and conscious abilities is no longer linked to the ontological distinction between, on the one hand, a reality that is in itself inexpressible and, on the other hand, a reality that is in principle expressible (cf. Buzzoni and Savojardo 2019 and Savojardo 2013).

able to identify an object or perform an activity. By this route Polanyi arrives at the fundamental conclusion that there are two views, two ways of being aware of the same reality: a 'subsidiary' or 'tacit' awareness of the details, which allows us, at a deep level, to grasp the object in its entirety, and a 'focal' or direct awareness of the details, in which the comprehensive unity tends to dissolve into a myriad of details (cf. Polanyi 1969: 113-14).

These two types of views, intentionality or awareness, express the non-ontological but functional way (what is focal can become subsidiary, or vice versa), in which Polanyi draws the distinction between explicit and tacit knowledge, a way that is decisive for the issue of understanding the intelligent use of the tool (cf. Buzzoni and Savojardo 2019). Polanyi himself illustrates the distinction between focal and subsidiary awareness with the example of using a hammer: while we use a hammer to drive a nail into the wall, we pay attention to both objects, but in an entirely different way. We try, in fact, to use the hammer in a certain way, mindful of the blows on the nail: we are primarily interested in achieving our goal, but "we are certainly alert to the feelings in our palm and the fingers that hold the hammer. They guide us in handling it effectively" (Polanyi 1958: 57). It is evident that the use of the instrument cannot be separated from that of our own body, to which the same distinction between subsidiary and focal awareness can be applied. The fact that all our conscious interventions in reality involve the subsidiary use of our bodies means that this can be defined as "the only aggregate of things of which we are aware almost exclusively in such a subsidiary manner" (Polanyi 1969: 214).

When we learn to use a new tool or when we use an already known tool in a new way, it is as if we extend our bodily equipment to include the tools we have encountered. The tool becomes part of our bodily system and the mind relates to it as an element of its own body, and thus as a part of itself as an entity acting in the world.

The knowledge of our body, like that of the tools we assimilate to it, when we intervene in a technical manner on the reality, in most cases, is a knowledge that remains at a tacit level. Tacit knowledge is, in fact, repeatedly defined by Polanyi as unlimited knowledge through which we tacitly understand something about ourselves as persons engaged in the search for truth. It is an implicit knowledge that concerns the indirect or 'subsidiary' awareness of ourselves, of the skills and tools that we assimilate into our personal being: "*We always know tacitly that we are holding of our explicit knowledge to be true*" (Polanyi 1959: 12). That 'we' includes our being living bodies in a space of action that is only part of the cultural reality in which we have always been embedded. Everything that relates the person to the context that surrounds him or her has an instrumental value starting from the body, from the tools we use in our daily lives, up to the most complex information technologies. From this point of view, words and concepts also have a similar instrumental value connected to the person who uses them to make explicit knowledge that was initially only implicit and to communicate. In an interesting passage, for example, Polanyi (1969: 145) constructs a parallelism between the acquisition of a language and the use of a common tool, such as a stick: the transformation of meaningless sounds into words depends on the process of language acquisition, through which direct attention to sounds becomes attention from them, towards the object of reference. This vector property of language, linked to the principle of transparency, concerns those who master a language. The same can be said of the use of a stick to learn to move in the dark: when we

first use it we will pay attention to every blow against the palm and fingers of our hand, every time the stick encounters an object; but when we have learnt to use it in the correct manner we will no longer pay attention to the insignificant blows on our hand, but will pay attention from them to the end of the stick that intercepts the obstacles in the room. Words and instruments are such for me and thus become part of my personal (and not subjective) instrumental apparatus.

Through the Polanyian proposal of an epistemology of the human person, one can reconsider the distinction between motor and cognitive elements of knowing as a functional distinction, a difference in perspective, dependent on personal choice. The subsidiary awareness of my body, understood not so much as an object among others, but first and foremost as a lived body in action, accompanies all my verbal, conceptual or explicit knowledge. But if I wanted, I could at any time shift my focal attention to the subsidiary elements that make up my body, thus making the individual bodily organs the object of study and interest. And this, after all, is also what the surgeon does while operating: he does not see the organs as subsidiary elements of a living body, of a person embedded in his or her environment, but regards them directly as individual objects worthy of attention in themselves. The change of perspective on the real, however, cannot be understood except by referring to that place of personal encounter between subsidiary and focal elements of knowledge, that centre of commitments and interests that is the human person. From this point of view, the intelligent use of any tool, from the stick for moving in the dark, to the terms of one's own language, to the technical instruments of a specific scientific discipline, becomes the use of a piece of nature in a personal project, connected to the space of action of a body understood first and foremost as that set "of things known almost exclusively by relying on our awareness of them for attending to something else" (Polanyi 1969: 147).

Always, when we intervene in a technical-practical sense on empirical reality, we do so by using different tools (conceptual and otherwise) that affect us and are part of us as persons. Consider, for example, the quote at the beginning of this paper, from which it emerges that the ability to paint and see things in a certain way, for Renoir (paralysed by arthritis), no longer resided in the individual co-body organs, it had 'shifted' to the instrument which became part of the person as a body-mind unit. This description clearly exemplifies how an instrument becomes an integral part of the person as an inseparable body-mind unit, whose body is capable of intervening in reality because it is guided by a type of reasoning that is never, from an empirical point of view, 'pure' or separated from the sensorimotor sphere. This statement, however, does not imply any reductionism of the mental to the physical, let alone a form of philosophical dualism. The way we represent reality in our technical-practical intervention in it is determined by the type of perspective we decide to put into practice in our 'attempts' at problem solving, ranging from solving simple problems in everyday life to studying complex and intricate situations in the natural and social sciences. Consciously, we can, in fact, decide to direct our focal attention to all the clues or details that are part of us because they are part of our 'subsidiary' equipment by means of which we deal with different problem situations (but in this way we will lose an overall view); or we can choose to look from these subsidiary aspects and beyond them to grasp the solution to the problem, in a unitary sense.

As already pointed out, in the personal being understood as an inseparable unity of mind-body, the mind can be aware of body parts, as well as of all those instruments (conceptual and otherwise) that are integrated into our person in a

direct (focal) or indirect (subsidiary) manner, and these two types of awareness also generate two ways of understanding the body-mind relationship: if we consider the individual organs in themselves, these become objects among others and the activity of consciousness is lost sight of; if, on the other hand, we consider the bodily mechanisms as subsidiary elements on which the mind relies in its conscious activities, the individual organs take on a new meaning in the inseparable mind-body unity always included in a certain space of action.

This type of functional 'mechanism', in the intelligent use of any tool, concerns both common knowledge and scientific knowledge, since every cognitive pathway develops in the interweaving of tacit and explicit, of subsidiary and focal elements, of corporeal and conceptual elements. For this reason, we have argued there is no practical, tacit or sense-motor knowledge, exclusively connected to the body, separate from another explicit or conceptual cognitive sphere: the distinction between so-called common knowledge and scientific knowledge cannot be a distinction of principle that presupposes, on an ontological level, two separate cognitive contexts. However, if we think about interchangeability relation between subsidiary and focal elements of knowledge described by Polanyi, we can reconsider the distinction between corporeal and cognitive elements of knowledge, with reference to the two different perspectives that the human person, embedded in a certain cultural, linguistic, social context, can choose to assume. In what does science consist if not in the attempt to translate the tacit into the explicit, through experiment? Although this 'translation' work takes place all the time also in common knowledge, it is stronger and more evident in science, where it is often very arduous and may take several years, than in common knowledge, for which we almost never feel the need to focus on the subsidiary elements that enable us to perform certain activities, such as walking, swimming or cycling, despite the fact that this possibility is always contemplated. From this point of view, the only difference between common knowledge and scientific knowledge can only be a difference of degree, and not of principle, since science, while developing at a more elaborate level, already contains and is always nourished by common knowledge, through a series of tacit skills that bind us to one another, in a universe that takes on the character of the person.

4. Conclusion

How do we represent reality when we act on it in a technical-practical sense, through the intelligent use of particular tools?

The paper attempted to answer this initial question by analysing the relationship between sensorimotor and cognitive aspects in the intelligent use of tools, a relationship that shows a continuity between common knowledge and scientific knowledge. The problem of the intelligent use of tools can thus be considered as a paradigmatic case useful in highlighting the link between these cognitive domains, the difference between which cannot be a difference in principle, but only in degree. The conclusion we have reached is supported by a series of arguments developed in the first and second sections respectively.

The first part of the paper framed the problem of the intelligent use of tools from the perspective of Embodied Cognition, in order to highlight some important philosophical issues that emerge in the light of the experimental neuroscientific results. With reference to Embodied Cognition, in fact, two different trends have arisen: on the one hand, the tendency to claim that tool use depends

exclusively on the action of the sensorimotor system; on the other hand, the tendency to describe a type of technical reasoning or mechanical problem solving, separate from the sensorimotor system. With this problematic situation in mind, an attempt has been made to highlight certain philosophical assumptions implicit in the neuroscientific debate. The separation between cognitive and motor aspects in the intelligent use of tools, if absolutized, risks becoming a difference between common, practical knowledge and scientific, abstract knowledge.

In the second part of this paper, an attempt was made to demonstrate that this difference in principle is not sustainable in our representation of reality, mediated by our technical, practical, instrumental intervention in it. Our technical intervention in reality is always mediated by concept, and our reasoning is always, to a certain extent, connected to practical action; the use of any instrument, therefore, in our field of action, is always intelligent, conceptually mediated use. This, in principle, applies both to the more mundane tools we use in our everyday lives and, at a more elaborate level, to the construction and use of experimental machines in the various scientific and technological practices. 'Pure' thought and action cannot be separated on the level of experimental science: our very reasoning in a counterfactual manner ends up being the condition of possibility of our intervention in reality (cf. above all Buzzoni 2008), both in common thought and, on a more elaborate level, in science.

In the intelligent use of tools, in any context, from the simplest and most immediate to the most complex, cognitive and motor elements of knowledge are always intertwined. The distinction in principle between common knowledge and scientific knowledge loses its meaning. However, in the last part of the text, I argued for a new way of understanding the relationship between the motor and cognitive elements of knowing, a way through which a difference of degree, and not of principle, between common and scientific thinking emerged. To this end, the reference to the functional mechanism described by M. Polanyi and founded on the distinction between subsidiary elements, mostly connected to the dimension of one's own body, and focal elements of knowing, which consist essentially in the conceptual formulation of a tacit knowledge that moves within and with our personal being, was useful. The reference to the Polanyian epistemology of the person has allowed us to consider the distinction between sensorimotor and cognitive aspects of knowing in a functional rather than ontological sense. It is up to the person to choose to move from a subsidiary awareness of those elements that are part of us and include, along with our body, the tools we assimilate to it, to a focal or direct awareness of them. The shift is always, in principle, possible, since it is not a question of overcoming the leap between two different, separate spheres, from an ontological point of view, but only of a change of outlook, functional to the context and situation in which the person is placed, in his or her daily activities, as in science. The personal and conscious decision to shift from one perspective to the other concerns all knowledge, even though, such a shift from the tacit to the explicit, or vice versa, is a fundamental requirement in the experimental sciences, rather than in everyday problem-solving.

The answer to the initial question on the representation of reality when we intervene on it through the intelligent use of tools highlighted the need to hold together the cognitive and motor elements of action-driven knowing. This need highlights a link between the plane of common knowledge and that of scientific knowledge. The only difference between these can only be a difference of degree, since when we represent reality, using certain tools in our daily practices, we

hardly ever feel the need to change our perspective of analysis, passing, according to Polanyi's language, from a subsidiary view to a focal view of the particulars of that activity; the issue is quite different, however, for science, whose primary aspiration is to translate the tacit into the explicit as much as possible, in order to arrive at a knowledge that can be reconstructed by the entire community.[6]

References

Bennett, M., Dennett, D., Hacker, P. and Searle, J. 2007, *Neuroscience and Philosophy: Brain, Mind, and Language*, New York: Columbia University Press.

Bennett, M.R. and Hacker, P.M.S. 2022, *Philosophical Foundations of Neuroscience,* 2nd edition, Hoboken: Wiley Blackwell.

Berlucchi, G. and Aglioti, S.M. 2010, "The Body in the Brain Revisited", *Experimental Brain Research*, 200, 1, 25-35.

Bianchi, I. and Savardi, U. 2006, "Oppositeness in Visually Perceived Forms", *Gestalt Theory*, 4, 354-74.

Bianchi, I., Bertamini, M., Burro, R., and Savardi, U. 2017a, "Opposition and Identicalness: Two Basic Components of Adults' Perception and Mental Representation of Symmetry", *Symmetry*, 9, 128, DOI: 10.3390/sym9080128

Bianchi, I., Paradis, C., Burro, R., Van de Weijer, J., Nyström, M., and Savardi, U. 2017b, "Identification of Opposites and Intermediates by Eye and by Hand", *Acta Psychol.*, 180, 175-89, DOI: 10.1016/j.actpsy.2017.08.011

Bianchi, I., Branchini, E., Burro, R., Capitani, E., and Savardi, U. 2020, "Overtly Prompting People to "Think in Opposites" Supports Insight Problem Solving", *Thinking & Reasoning*, 26, 31-67, DOI: 10.1080/13546783.2018.1553738

Branchini, E., Bianchi, I., Burro, R., Capitani, E., and Savardi, U. 2016, "Can Contraries Prompt Intuition in Insight Problem Solving?", *Front. Psychol.*, 7, DOI: 10.3389/fpsyg.2016.01962

Branchini, E., Capitani, E., Burro, R., Savardi U., and Bianchi, I. 2021, "Opposites in Reasoning Processes: Do We Use Them More Than We Think, But Less Than We Could?", *Frontiers in Psychology*, 12, DOI: 10.3389/fpsyg.2021.715696

Brandi, M.L., Wohlschäger, A., Sorg, C., and Hermsdörfer, J. 2014, "The Neural Correlates of Planning and Executing Actual Tool Use", *Journal of Neuroscience*, 34, 39, 13183-194.

Buzzoni, M. 1995, *Scienza e tecnica: Teoria ed esperienza nelle scienze della natura*, Roma: Studium.

Buzzoni, M. 2004, *Esperimento ed esperimento mentale*, Milano: Angeli.

Buzzoni, M. 2005, "Scienza e tecnica", *Dialoghi*, 5, 3, 20-25.

Buzzoni, M. 2008, *Thought Experiment in the Natural Sciences: A Transcendental-Operational Conception*, Würzburg: Königshausen & Neumann.

Buzzoni, M. and Savojardo, V. 2019, "L'uso intelligente dello strumento fra embodied cognition e teoria polanyiana della conoscenza tacita", in Allegra, A.,

[6] This work was supported by the Italian Ministry of University and Research through the PRIN 2017 project "The Manifest Image and the Scientific Image" prot. 2017ZNW W7F_004.

Calemi, F., and Moschini, M. (a cura di), Alla *fontana di Siloe: Studi in onore di Carlo Vinti*, Napoli-Salerno: Orthotes, 151-65.

Byrne, R.M.J. 2016, "Counterfactual Thought", *Annu. Rev. Psychol.*, 67, 135-57, DOI: 10.1146/annurev-psych-122414-033249

Byrne, R.M.J. 2018, "Counterfactual Reasoning and Imagination", in Ball, L.J. and Thompson, V.A. (eds.), *The Routledge Handbook of Thinking and Reasoning*, London: Routledge, 71-87.

Cardinali, L., Frassinetti, F., Brozzoli, C., Urquizar, C., Roy, A.C., and Farnè, A. 2009, "Tool-Use Induces Morphological Updating of the Body Schema", *Current Biology*, 19, R478-R479.

Caruana, F. 2012, "Strumenti intelligenti: Che cosa accade nel cervello quando estendiamo il corpo", *Sistemi intelligenti*, 1, 127-39.

Caruana, F. and Cuccio, V. 2015, "Il Corpo come icona: Abduzione, strumenti e Embodied Simulation", *Quaderni di Studi Semiotici*, 120, 93-101.

Chao, L.L. and Martin, A. 2000, "Representation of Manipulable Man-Made Objects in the Dorsal Stream", *NeuroImage*, 12, 4, 478-84.

Clark, A. and Chalmers, D. 1998, "The Extended Mind", *Analysis*, 58, 10-23.

Dumas, D., Alexander, P.A., and Grossnickle, E.M. 2013, "Relational Reasoning and Its Manifestations in the Educational Context: A Systematic Review of the Literature", *Educ. Psychol. Rev.*, 25, 391-42, DOI: 10.1007/s10648-013-9224-4

Evans, J.S.B.T. 2007, *Hypothetical Thinking: Dual Processes in Reasoning and Judgment*, Hove: Psychology Press.

Ferretti, G. 2021, "A Distinction Concerning Vision-for-Action and Affordance Perception", *Conscious Cogn.*, DOI: 10.1016/j.concog.2020.103028

Gibson, J.J. 1979, *The Ecological Approach to Visual Perception*, Boston: Houghton Mifflin.

Goldenberg, G. 2013, *Apraxia: The Cognitive Side of Motor Control*, Oxford: Oxford University Press.

Goldenberg, G. and Hagman, S. 1998, "Tool Use and Mechanical Problem Solving in Apraxia", *Neuropsychologia*, 36, 581-89.

Goldenberg, G. and Spatt, J. 2009, "The Neural Basis of Tool Use", *Brain,* 132, 1645-55.

Grafton, S.T., Fadiga, L. Arbib, M.A., and Rizzolatti, G. 1997, "Premotor Cortex Activation during Observation and naming of Familiar Tools", *NeuroImage*, 6, 231-36.

Hegarty, M. 2004, "Mechanical Reasoning by Mental Simulation", *Trends in Cognitive Sciences*, 8, 280-85.

Hutchins, E., 1995, *Cognition in the Wild*, Cambridge, MA: MIT Press.

Iriki, A., Tanaka, M., and Iwamura, Y. 1996, "Coding of Modified Body Schema During Tool Use by Macaque Postcentral Neurones", *Neuroreport*, 7, 14, 2325-30.

Johnson-Frey, S.H. 2003, "What's So Special about Human Tool Use?", *Neuron*, 39, 2, 201-204.

Makris, S., Hadar, A.A., and Kielan, S. 2013, "Are Object Affordances Fully Automatic? A Case of Covert Attention", *Behavioral Neuroscience*, 127, 5, 797-802.

Maravita, A. and Iriki, A. 2004, "Tools for the Body (Schema)", *Trends in Cognitive Sciences*, 8, 2, 79-86.

Menary, R. 2010, *The Extended Mind*, Cambridge, MA: MIT Press.

Noë, A. 2004, *Action in Perception*, Cambridge, MA: MIT Press.

Osiurak, F, Jarry, C., Aubin, G., Allain, P., and Etcharry-Bouyx, R. 2009, "Unusual Use of Objects after Unilateral Brain Damage: The Technical Reasoning Model", *Cortex*, 45, 769-83.

Osiurak, F., Jarry, C., and Le Galle, D. 2010, "Grasping the Affordances, Understanding the Reasoning: Toward a Dialectical Theory of Human Tool Use", *Psychological Review,* 117, 2, 517-40.

Osiurak, F. 2014, "What Neuropsychology Tells Us about Human Tool Use? The Four Constraints Theory: Mechanics, Space, Time, and Effort", *Neuropsychology Review,* 24, 2, 88-115.

Palmiero, M. and Borsellino, M.C. 2018, *Embodied Cognition: Comprendere la mente incarnata*, 2nd edition, Fano: Aras.

Polanyi, M. 1958, *Personal Knowledge: Towards a Post-Critical Philosophy*, London: Routledge and Kegan Paul; quotes from II edit. 1962.

Polanyi, M. 1959, *The Study of Man*, London: Routledge and Kegan Paul; quotes from edit. 2014 Mansfield Centre: Martino Publishing.

Polanyi, M. 1969, *Knowing and Being*, London: Routledge & Kegan Paul.

Rupert, R.D. 2009, *Cognitive Systems and the Extended Mind*, New York: Oxford University Press.

Sakreida, K. et al. 2016, "Affordance Processing in Segregated Parieto-Frontal Dorsal Stream Sub-Pathways", *Neuroscience and Biobehavioral Reviews*, 69, 89-112, DOI: 10.1016/j.neubiorev.2016.07.032

Sartori, L., Straulino, E., and Castiello, U. 2011, "How Objects Are Grasped: The Interplay between Affordances and End-Goals", *PLoS ONE*, 6, 9, DOI: 10.1371/journal.pone.0025203

Savojardo, V. 2013, *Scienza, fede e verità personale in Michael Polanyi*, Roma: Aracne.

Shapiro, L. 2010, "Embodied Cognition", in Margolis, E., Samuels, R., and Stich, S. (eds.), *Oxford Handbook of Philosophy and Cognitive Science*, Oxford: Oxford University Press.

Shapiro, L. 2019, *Embodied Cognition*, 2nd edition, New York: Routledge.

Thompson, E. 2007, *Mind in Life*, Cambridge, MA: Harvard University Press.

Valyeaar, K.F., Cavina-Pratesi, C., Stiglick, A.J., and Culham, J.C. 2007, "Does Tool-Related fMRI Activity within the Intraparietal Sulcus Reflect the Plan to Grasp?", *NeuroImage*, 36, 2, T94-T108.

Varela, F., Thompson, E., and Rosch, E. 1991, *The Embodied Mind*, Cambridge, MA: MIT Press.

Williamson, T. 2016, "Knowing by Imagining", in Kind, A. and Kung, P. (eds.), *Knowledge Through Imagination*, New York: Oxford University Press, 113-23.

Williamson, T. 2020, "Book Review: Arnon Levy, Peter Godfrey-Smith (eds.), The Scientific Imagination: Philosophical and Psychological Perspectives", *Notre Dame Philosophical Reviews*, https://ndpr.nd.edu/reviews/the-scientific-imagination-philosophical-and-psychological-perspectives/.

Wilson, M. 2004*, Boundaries of the Mind: The Individual in the Fragile Sciences: Cognition,* Cambridge: Cambridge University Press.

Does Evolution Favor Accurate Perception?[1]

Adriano Angelucci, Vincenzo Fano,* Gabriele Ferretti,** Roberto Macrelli,* and Gino Tarozzi**

** University of Urbino Carlo Bo*

*** Ruhr University Bochum*

Abstract

The currently mainstream view is that, in normal conditions, our perceptual representations are largely accurate, as natural selection tends to favor epistemically reliable perceptual systems. This latter assumption has been questioned by Donald Hoffman and his collaborators by drawing on the formal tools of evolutionary game theory. According to their model, an organism whose visual system were tuned to objective reality would be driven to extinction. We argue that their model fails to take environmental modifications into due account, and we show that, once such changes are incorporated into the model, the latter will predict that an organism whose visual representations are at least partially accurate will in fact be more successful from an evolutionary point of view.

Keywords: Perception, Perceptual strategies, Evolutionary game theory.

1. Introduction

The currently mainstream view among scientists studying perception is that, in normal conditions, our perceptual representations are largely accurate—i.e., that, to some extent, they do a good job at tracking the objective structure of the external world.[2] The view in question usually rests on a specific evolutionary assumption—i.e., that natural selection will in the long run favor individuals whose perceptual systems are epistemically reliable. Within the relevant literature it is indeed typically argued that if our perceptual representations were not somehow tuned to the objective structure of reality, evolutionary pressures would long have driven our species to extinction.[3] In a series of papers, Donald Hoffman and his

[1] In this article we bring out what we take to be the main philosophical consequences of the two models presented in Angelucci et al. 2021.

[2] Cf., e.g., Marr 1982: 340, Trivers 2011: 2, and Pizlo et al. 2014: 227.

[3] Cf., e.g., Geisler & Diehl 2003, and Yuille & Bülthoff 1996.

Argumenta 9, 1 (2023): 105-114
ISSN 2465-2334

© 2023 The Authors

DOI 10.14275/2465-2334/202317.ang

collaborators (henceforth, H&C) made use of evolutionary game theory in order to question this widely held assumption.[4] Evolutionary games, in their view, would conclusively establish that our visual systems are in fact tuned to *utility*, not to objective reality. As a consequence, H&C maintain, we would have little or no reason to believe that our visual representations are always, or even usually, accurate.

The far-reaching philosophical implications of this purported fact about human vision can hardly be overstated. Arguably, if H&C's conclusion were to prove correct, then large swaths of contemporary epistemology and philosophy of mind would have to be called into question, to say the least. Naturalistic approaches to knowledge and justification, for instance, are more or less explicitly premised on the assumption that, in normal conditions, our perceptual systems and processes are generally reliable,[5] and the same seems to hold for naturalistically minded accounts of the semantic content of our mental states.[6] Moreover, H&C's conclusion, if true, would arguably lend significant support to the skeptical—yet nonetheless popular in some intellectual milieus—idea according to which empirical science would not in the end be entitled to any justified claims about what the external world is like, independently of the way in which it happens to be perceived or thought of by sentient beings.

In what follows, we intend to argue that H&C's epistemically grim conclusion is still far from being the only one licensed by the formal tools of evolutionary game theory. Our main goal will be to show that, contrary to their view, the mere fact that the complex evolutionary dynamics responsible for shaping our perceptual systems will in the long run increase our fitness does not entail that our visual representations will therefore be generally inaccurate. What led H&C astray, in our view, is that their model fails to take the relevance of environmental modifications into due account. As we will try to show, however, a model that incorporates a dynamic, rather than static view of the organism's environment, will predict that—up to a certain point—the acquisition of apparently useless information about said environment will in fact increase fitness.[7] In particular, we suggest that this will be the case even when the organism which detects such apparently useless information and the one which does not make use of the same number of bits. Our model then suggests that, in general, an organism whose visual representations were at least partially accurate would be more successful from an evolutionary point of view.

[4] Cf., in particular, Mark, Marion, and Hoffman 2010, Hoffman and Manish 2012, Hoffman, Manish, and Mark 2013, Hoffman, Manish, and Prakash 2015.

[5] Alvin Goldman, the father of *process reliabilism*, found it plausible to suppose that "many cognitive functions subserving the attainment of true beliefs […] were selected for in evolution because of their biological consequences, that is their contribution to genetic fitness" (Goldman 1986: 98, quoted in Stich 1990: 161).

[6] Consider, e.g., the following two passages from Ruth Millikan and Daniel Dennett respectively: "The mechanisms in us that produce beliefs […] all have in common at least one proper function: helping to produce true beliefs" (Millikan 1984: 317, quoted in Stich 1990: 162); "natural selection guarantees that *most* of an organism's beliefs will be true" (Dennett 1981: 75, quoted in Stich 1990: 55).

[7] As we shall see, the information in question is here said to be 'apparently' (as opposed to 'actually') useless in the sense that, by gathering it, the organism will incur costs which—while increasing its fitness *in the long run*—are bound to have an *immediate* negative impact in terms of fitness. Thanks to an anonymous referee for inviting us to clarify this point.

The plan is as follow. In the next section we will introduce a basic formal framework which allows us to define three different perceptual strategies, dubbed *realist*, *critical realist* and *interface* strategy respectively. This framework will also provide us with the means to gauge the accuracy of each strategy. In section 3 we will then consider and assess H&C's argument for the inaccuracy of our visual representations, according to whose general conclusion an interface strategy will in the long run clearly outcompete a critical realist one. In section 4 we put forward an alternative model in order to show that, once a biologically more realistic view of the environment is incorporated into the model, a critical realist strategy will in the long run outcompete an interface one. In section 5 we will sum up our considerations and draw some conclusions.

2. Perceptual Strategies

In line with most contemporary philosophical theories of perception we will assume that perception is at bottom a representational process, i.e., that our perceptual systems represent reality by ascribing various features to individual objects as well as to the visual scene as a whole.[8] As H&C focus on vision, our first task will consist in developing a plausible and empirically testable model of visual perception, accordingly conceived as a process whereby a given environmental stimulus causally interacts with our visual system, thereby giving rise to a more or less accurate representation of its source—i.e., a *visual representation*. So let us do just that.

Our model—just as any model—will inevitably involve a fair amount of idealization. So let us begin by thinking of an organism's environment as a given set **E** of features. Every subset of **E** can then be seen as a stimulus capable of causing in the organism a corresponding subset of a further set **V** of visual representations. Let us now call $\mathbf{T_E}$ and $\mathbf{T_V}$ the "best possible theories" of, respectively, **E** and **V**, and let us further conveniently suppose that these two theories are developed enough to possess their respective state-spaces $\mathbf{S_{TE}}$ and $\mathbf{S_{TV}}$.[9] By so doing, we can then let a *representation function* **F** stand for the organism's ability to visually represent its environment, and an inverse *causal function* **Q** stand for the environment's causal effects on the organism's visual system—whereas **F** will map $\mathbf{S_{TV}}$ regions onto $\mathbf{S_{TE}}$ ones, **Q** will map $\mathbf{S_{TE}}$ regions onto $\mathbf{S_{TV}}$ ones. A *perceptual strategy*, at this point, will be a composite function **FQ** that maps $\mathbf{S_{TE}}$ regions onto $\mathbf{S_{TE}}$ ones.

We can now provide an exact definition of three distinct perceptual strategies that, following H&C, we may call *realist*, *critical realist*, and *interface* respectively.

[8] Cf., e.g., Nanay 2013, Siegel 2006, Brogaard 2014, and Ferretti & Zipoli Caiani 2019. In spite of various interesting attempts at developing nonrepresentational views of perception (cf., e.g., Noë 2004, Chemero 2009, and Hutto and Myin 2013), representationalism still remains the dominant view on the matter, and this is arguably mainly due to the undeniable explanatory advantages of the latter (cf. Pautz 2010, Nanay 2013), especially in case of the study of perceptual reality (Ferretti, forthcoming). It is however clear that, if perception were direct even in a weak sense, then H&C would be a fortiori wrong.

[9] We hasten to add that, for the purposes of the present argument, there is no need to think of our 'best possible theories' as actual scientific theories—T_E and T_V are rather intended as merely useful fictions whose sole purpose in what follows will be to illustrate our proposal concerning the measurement of visual representations' accuracy. Thanks to an anonymous referee for inviting us to clarify this point.

Letting $\mathbf{r_{SE}}$ stand for a given region of $\mathbf{S_{TE}}$, we can postulate that a perceptual strategy will be a *realist* one if $\mathbf{FQr_{SE}} = \mathbf{r_{SE}}$—i.e., if our visual representations perfectly mirror the environmental stimuli that give rise to them. A strategy will instead be a *critical realist* one if there is at least a subspace of $\mathbf{S_{TE}}$ (call it $\mathbf{S'_{TE}}$) within which, as it were, realism holds—i.e., within which, if $\mathbf{r_{S'E}}$ is a region of $\mathbf{S'_{TE}}$, and $\mathbf{S'_{TE}} \subset \mathbf{S_{TE}}$, then $\mathbf{FQ\ r_{S'E}} = \mathbf{r_{S'E}}$. A strategy will finally be an *interface* one if $\mathbf{S'_{TE}} = \varnothing$[10] (cf. Fig. 1).

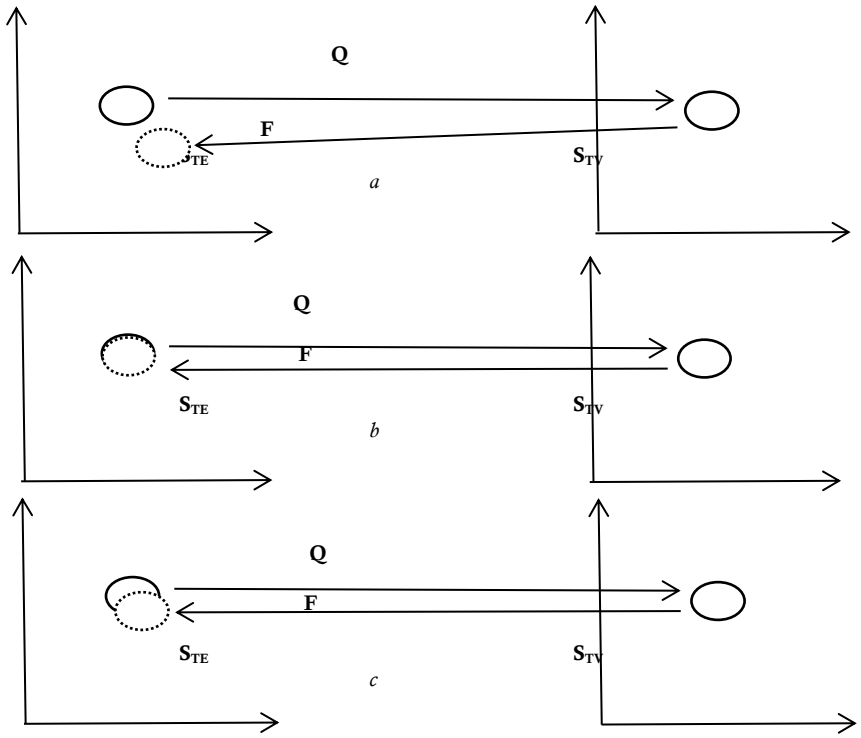


Fig. 1 – A certain set of stimuli—represented by the ellipse in the state space of the environment ($\mathbf{S_{TE}}$)—causes the changes described by function **Q**, i.e., a certain modification in the visual field of the organism ($\mathbf{S_{TV}}$)—the ellipse on the right. Such modification in turn constitutes an attempt to represent (**F**) the initial set of stimuli—the dotted ellipse on the left. There will hence be three possible situations: *a*: Interface strategy, *b*: Realist, and *c*: Critical Realist.

In light of the above, we can now think of the accuracy of our visual representations as a correspondence between the two state-spaces $\mathbf{S_{TE}}$ and $\mathbf{S_{TV}}$. In particular, our framework will allow us to measure such accuracy through the *distance* $\mathbf{d_{EV}}$ between the (objective) conjunct probability measure on $\mathbf{S_{TE}}$—i.e., $\boldsymbol{\mu}_{\mathbf{EE}}$—and the conjunct probability measure on $\mathbf{S_{TV}}$—i.e., $\boldsymbol{\mu}_{\mathbf{VV}}$. This last point perhaps requires

[10] However, an organism implementing an interface strategy will still be sensible to *environmental discontinuities*, and it will therefore preserve a residual representational capacity. Our definition is only meant to capture the idea that the representational contents of a perceptual system implementing such a strategy will be so far removed from a completely accurate representation of the environment as to have virtually zero accuracy. Thanks to an anonymous referee for inviting us to clarify this point.

some clarification. Perceptual strategies have earlier been defined relative to the state-space of our "best possible theory" $\mathbf{T_E}$ of the environment—i.e., $\mathbf{S_{TE}}$. It must be kept in mind, however, that neither an individual visual representation belonging to $\mathbf{S_{TV}}$, nor an individual stimulus belonging to $\mathbf{S_{TE}}$ are themselves directly accessible for us. As a consequence, the only viable way to assess the relevant distance ($\mathbf{d_{EV}}$)—and to thereby decide which one of the three perceptual strategies is actually being implemented—will be to rely on the conjunct probabilities of distinct stimuli, and of distinct visual representations respectively. In operational terms, then, the relevant question will not have the form: What is the probability that the organism will experience a visual representation of red, given that a red distal stimulus is being instantiated? But rather: What is the probability that it will experience two adjacent visual representations (e.g., a green and a red one), given that two corresponding adjacent distal stimuli—a green and a red one—are being instantiated? It is the answer to this latter question that will in fact give us a measure of the visual system's accuracy.

The distance $\mathbf{d_{EV}}$ can then be normalized so that, when two measures are the same, its value will be "0", and when two given elements $\mathbf{x_E}$ and $\mathbf{y_E}$ of an algebra defined on $\mathbf{S_{TE}}$ are such that "μ_{VV} (x_V, y_V) = 1 – μ_{EE} (x_E, y_E)" its value will be "1". At this point it will be reasonable to posit that a critical realist strategy will determine a value of $d_{EV} \leq 0.5$, an interface strategy will determine a value of $d_{EV} > 0.5$,[11] and a realist strategy will hold when $d_{EV} = 0$. With this formalism in place, let us now move on to consider H&C's main argument for the purported inaccuracy of our visual representations by focusing on the interplay amongst the perceptual strategies defined above.

3. The Case for Interface

Evolutionary game theory is arguably the best way to predict the evolution of a discrete phenotypic trait whose fitness depends on its frequency within a population.[12] The general idea is that a trait's fitness could be *affected* by its frequency. Consider, for instance, the random appearance, on a butterfly's wing, of a pigmented region which just so happens to mimic the eye of a snake. This random mutation will presumably have the immediate effect of decreasing the butterfly's chances to be eaten by a bird, thereby increasing its fitness. The mutation in question, however, will only have this effect (i.e., misleading birds into believing that a butterfly is a snake) if it makes its appearance in a limited number of butterflies.[13] In our present case, the trait will of course be a perceptual strategy coexisting with other strategies, and whose fitness will therefore also depend on the frequency of its rivals. As we anticipated above, H&C hold that fitness-maximization is bound to have a negative impact on the overall accuracy of our visual representations, as an interface strategy, in their view, would clearly outcompete—and hence, in the long run, drive to extinction—a critical realist one.[14] In order to substantiate

[11] For the sake of simplicity, we are here focusing on *binary* features only (such as, e.g., black/white). With respect to such features, it seems reasonable to assume that getting them right 50% of the time is tantamount to having zero information about the environment.

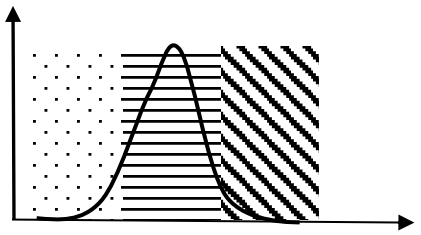
[12] Cf. Rice 2004: 263.

[13] We would like to thank an anonymous referee for encouraging us to clarify this point.

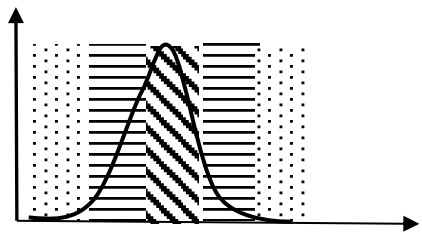
[14] Cf., e.g., Mark et al. 2010: 504.

this claim, they ask us to consider the following evolutionary game in which all three of our perceptual strategies—i.e., *realism*, *critical realism*, and *interface*—compete with each other.

The playing field features three different territories, and only one resource whose values range from 1 to 100. Utility—which is proportional to fitness—is represented by a Gaussian with its peak at 50, and it is therefore not proportional to the quantity of resource to be found on each territory. Now, whereas the *realist* strategy will gather all of the available information, the *critical realist* one will instead only rely on three visual representations (e.g., three different colors standing for different resource quantities), and the same will be the case for the *interface* strategy. The difference between the two latter strategies lies in the way in which the three colors are used (cf. Fig. 2).



Critical realist strategy



Interface strategy

Fig. 2 – Critical realist strategy (above); Interface strategy (below). On the x-axis the quantity of resource; on the y-axis the utility. The difference between critical realist and interface strategies is expressed through diverse distributions of colors. The different colours are represented through the diverse types of filling: points, horizontal lines and diagonal lines. It is evident that the use of colors in the interface strategy is more useful—in terms of fitness—than its counterpart in the critical realist strategy.

As we can see from the two graphs in Fig. 2, while the critical realist strategy will disregard the utility curve and simply associate the three colors with the increasing quantity of the resource, the interface strategy will keep track of utilities only. Now, as resource quantity and utility are non-monotonically related, each strategy will incur the costs associated with the process of gathering information about the environment and calculating its corresponding utility. It follows that the interface strategy will soon outcompete the critical realist one.

On closer inspection, however, this stage seems clearly and intentionally set to put critical realism at a disadvantage. Indeed, by keeping perceptual complexity fixed, the interface strategy will obviously have a running start. And yet, as we shall presently see, additional considerations may easily turn the tables on the

interface strategy. As it has been objected, for instance,[15] organisms often tend to homeostasis, and hence have an interest in knowing whether the quantity of a given resource happens to be above or below a certain threshold. When this is the case, the critical realist strategy will have an advantage over its interface counterpart. While this objection seems to point in the right direction, in the next section we will argue that H&C's view of human perception is beset by a more fundamental limitation of their model.

4. Making Room for Change

To shed light on what we regard as the main drawback of H&C's model, let us now consider the following simple case. Suppose that the organisms competing in our game are sparrows, and the resource are little worms. Given that worms evolve, we can easily imagine that a small and apparently inconsequential random mutation will at some point significantly decrease the size of a small number of individuals in their population. And we can further imagine that—as the sparrows' foraging strategy tends to zero in on bigger preys—the new trait will spread rapidly across the worms' population. This latter fact will in turn obviously alter the ratio between the utility of the resource and its quantity (expressed in number of worms). The point now is that, under the imagined circumstances, a sparrow implementing an interface strategy will accordingly still "think" that the same number of worms is needed in order to maximize utility, and will hence end up lagging behind in terms of fitness.[16] Its critical realist competitor, on the other hand, will "know better" and accordingly move to an area where either more or bigger worms are to be found.

Cases similar to the above, we believe, clearly suggest that slight modifications in the environment can bring about serious disadvantages for organisms implementing an interface strategy. Indeed, by completely disregarding apparently useless information—such as, e.g., worms' size—the organisms in question will be utterly unresponsive to possible environmental modifications that do however have a significant impact on utility. Our main point is hence that, given a static environment, a strategy targeting utility will clearly outcompete one aimed at representing reality. In a situation where the environment changes, however, the opposite will be the case.

This can be shown by means of a very simple model in which an interface strategy will initially prevail over a critical realist one and yet this trend will reverse in due time because of modifications in the environment. According to the model in question, in other words, when the environment is held fixed and each

[15] Cf. Anderson 2015.

[16] The reason is that, immediately after the environmental change has taken place, a sparrow implementing an interface strategy will still lack the information that, in order to maximize utility, it will need to eat more worms. This is due to the fact that it will presumably take generations for a visual system implementing such a strategy to *retune* to the new utility distribution. As a real-life example of this dynamic, we can think of the extinction of dinosaurs after an asteroid hit the Yucatán Peninsula 66 million years ago thereby causing vast and sudden environmental changes. Their extinction was due to their incapacity to change rapidly their genetic code to face the new situation. Thanks to an anonymous referee for inviting us to clarify this point.

organism has the same number of bits at its disposal, a strategy aimed at increasing utility will outcompete one aimed at representing reality. As the environment changes, however, the opposite will be the case.

Let us consider a genus $\boldsymbol{\omega}$, divided in two species $\boldsymbol{\omega_{IF}}$ and $\boldsymbol{\omega_{CR}}$—each implementing an interface, and a critical realist strategy respectively—and let us suppose that the environment within which the two species compete features two resources, $\boldsymbol{x}$ and $\boldsymbol{y}$, whose density happens to fluctuate over time with a slight difference of phase. Let us then additionally suppose that the overall resource utility is not proportional to the mere *sum* of $\boldsymbol{x}$ and $\boldsymbol{y}$'s density, but that it also depends on a further term related to the *difference* between their phases. Consider now the different ways in which $\boldsymbol{\omega_{IF}}$ and $\boldsymbol{\omega_{CR}}$ will respectively go about gathering information. Whereas $\boldsymbol{\omega_{IF}}$ will approach this task just by assessing the exact resource utility of an initial environmental situation, $\boldsymbol{\omega_{CR}}$ will instead at least keep an approximate track of the density fluctuation in the two resources. In can be shown that, in a similar setting, $\boldsymbol{\omega_{IF}}$ would initially outcompete $\boldsymbol{\omega_{CR}}$, as its perceptual strategy will for a while do a better job at tracking utility. As time goes by, however, $\boldsymbol{\omega_{CR}}$' rough estimate of $\boldsymbol{x}$ and $\boldsymbol{y}$'s density fluctuations—i.e., its relative responsiveness to environmental changes—will prove extremely valuable, as it will allow for a much better long-term assessment of their utility. At the end of the day, then, $\boldsymbol{\omega_{CR}}$ will be better off than $\boldsymbol{\omega_{IF}}$ from an evolutionary point of view.[17]

While this simple model is admittedly limited in scope, the assumptions upon which it rests seem quite reasonable. We take those assumptions to be the following:

(1) Environments change.
(2) Many environmental features display an oscillating pattern.
(3) Utility is not in general the mere sum of two such features.
(4) A constant utility function is not appropriate to represent utility in a changing environment.
(5) Knowledge of the environmental features' variation, while itself insufficient to locate the real utility function, nonetheless seems a reasonable starting point to assess utility in a changing environment.

5. Conclusions

If perception is the only way to acquire information about our environment and it turns out to be not even partially accurate, then investigating Homo sapiens and its environment would amount to merely inspecting our subjectivity. Yet modern science's moral and cognitive mission also consists of pursuing fallible and revisable attempts at formulating justified hypotheses about Homo sapiens, its origins and the world it inhabits. Many cultural milieus encourage the idea that empirical science cannot make any justified claims about the external world, independently of the way in which that world is perceived or thought of. If perception were completely inaccurate, this idea would be reinforced. We believe, however, that whether and to what extent human perception accurately represents the world is an epistemological matter which can be empirically investigated at least indirectly by using evolutionary mathematical models. We showed the limits of H&C's attempts at establishing the negative impact on fitness of an accurate representation

[17] Cf. Angelucci et al. 2021 for the mathematical derivation of this result.

of the world. Our model is clearly only a sketch at this stage, and it certainly requires further development. Indeed, we are confident that, given reasonable assumptions concerning what should count as an accurate perceptual representation, it should be possible to empirically investigate the comparative fitness of different perceptual strategies along the lines suggested by H&C. We also believe, however, that such investigation should carefully take into account modifications in the environment.[18]

References

Anderson, B.L. 2015, "Where Does Fitness Fit in Theories of Perception?", *Psychonomic Bulletin & Review* 22, 6, 1507-11.

Angelucci, A., Fano, V., Ferretti, G., Macrelli, R., and Tarozzi, G. 2021, "Evolutionary Dynamics and Accurate Perception: Critical Realism as an Empirically Testable Hypothesis", *Philosophia Scientiae*, 25: 157-78.

Brogaard, B. 2014 (ed.), *Does Perception Have Content?*, New York: Oxford University Press.

Chemero, A. 2009, *Radical Embodied Cognitive Science*, Cambridge, MA: MIT Press.

Ferretti, G. (forthcoming), "For an Epistemology of Stereopsis", *Review of Philosophy and Psychology*.

Ferretti, G. and Zipoli Caiani, S. 2019, "Between Vision and Action: Introduction to the Special Issue", *Synthese*, Special Issue, DOI: 10.1007/s11229-019-02518-w

Geisler, W.S. and Diehl, R.L. 2003, "A Bayesian Approach to the Evolution of Perceptual and Cognitive Systems", *Cognitive Science*, 27, 3, 379-402.

Hoffman, D.D. and Manish, S. 2012, "Computational Evolutionary Perception", *Perception*, 41, 9, 1073-91.

Hoffman, D.D., Singh, M., and Mark, J.T. 2013, "Does Evolution Favor True Perceptions", in Rogowitz, B.E., Pappas, T.N., and de Ridder, H. (eds.), *Proceedings of the SPIE 8651, Human Vision and Electronic Imaging*, XVIII, 865104, DOI: 10.1117/12.2011609

Hoffman, D.D., Manish, S., and Prakash, S. 2015, "The Interface Theory of Perception", *Psychonomic Bulletin & Review*, 22, 6, 1480-1506.

Hutto, D.D. and Myin, E. 2013, *Radicalizing Enactivism: Basic Minds without Content*, Cambridge, MA: MIT Press.

Mark, J.T., Marion, B.B., and Hoffman, D.D. 2010, "Natural Selection and Veridical Perceptions", *Journal of Theoretical Biology*, 266, 4, 504-15.

Marr, D. 1982, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, San Francisco: Freeman.

[18] We would like to thank the audience at the 3rd conference of the P.R.I.N. "The Manifest Image and the Scientific Image": *Models, Structures and Representation*, held in Urbino in June 2022, where the ideas contained in this article were first presented. We also thank three anonymous reviewers for helping us improve on an earlier version of this manuscript. We acknowledge support by the Italian Ministry of Education, University and Research through the PRIN 2017 project "The Manifest Image and the Scientific Image" prot. 2017ZNWW7F_004. Gabriele Ferretti also acknowledges support from a Humboldt Fellowship, hosted by Professor Albert Newen at the Institute for Philosophy II, Ruhr University Bochum, Germany.

Nanay, B. 2013, *Between Perception and Action*, Oxford: Oxford University Press.

Noë, A. 2004, *Action in Perception*, Cambridge, MA: MIT Press.

Nowak, M.A. 2006, *Evolutionary Dynamics: Exploring the Equations of Life*, Cambridge, MA: Belknap Press.

Pautz, A. 2010, "Why Explain Visual Experience in Terms of Content", in Nanay, B. (ed.), *Perceiving the World*, New York: Oxford University Press, 254-309.

Pizlo, Z., Li, Y., Sawada, T., and Steinman, R.M. 2014, *Making a Machine That Sees Like Us*, New York: Oxford University Press.

Rice, S.H. 2004, *Evolutionary Theory: Mathematical and Conceptual Foundations*, Sunderland: Sinauer.

Siegel, S. 2006, "Which Properties Are Represented in Perception?", in Gendler, T.S. and Hawthorne, J. (eds.), *Perceptual Experience*, Oxford: Oxford University Press, 481-503.

Stich, S.P. 1990, *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*, Cambridge, MA: MIT Press.

Trivers, R. 2011, *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*, New York: Basic Books.

Yuille, A. and Bülthoff, H.H. 1996, "Bayesian Decision Theory and Psychophysics", in Knill, D.C. and Richards, W. (eds.), *Perception as Bayesian Inference*, Cambridge: Cambridge University Press, 123-62.

Advisory Board

SIFA former Presidents

Eugenio Lecaldano (Roma "La Sapienza"), Paolo Parrini (University of Firenze), Diego Marconi (University of Torino), Rosaria Egidi (Roma Tre University), Eva Picardi (University of Bologna), Carlo Penco (University of Genova), Michele Di Francesco (IUSS), Andrea Bottani (University of Bergamo), Pierdaniele Giaretta (University of Padova), Mario De Caro (Roma Tre University), Simone Gozzano (University of L'Aquila), Carla Bagnoli (University of Modena and Reggio Emilia), Elisabetta Galeotti (University of Piemonte Orientale), Massimo Dell'Utri (University of Sassari), Cristina Meini (University of Piemonte Orientale)

SIFA charter members

Luigi Ferrajoli (Roma Tre University), Paolo Leonardi (University of Bologna), Marco Santambrogio (University of Parma), Vittorio Villa (University of Palermo), Gaetano Carcaterra (Roma "La Sapienza")

Robert Audi (University of Notre Dame), Michael Beaney (University of York), Akeel Bilgrami (Columbia University), Manuel Garcia-Carpintero (University of Barcelona), José Diez (University of Barcelona), Pascal Engel (EHESS Paris and University of Geneva), Susan Feagin (Temple University), Pieranna Garavaso (University of Minnesota, Morris), Christopher Hill (Brown University), Carl Hoefer (University of Barcelona), Paul Horwich (New York University), Christopher Hughes (King's College London), Pierre Jacob (Institut Jean Nicod), Kevin Mulligan (University of Genève), Gabriella Pigozzi (Université Paris-Dauphine), Stefano Predelli (University of Nottingham), François Recanati (Institut Jean Nicod), Connie Rosati (University of Arizona), Sarah Sawyer (University of Sussex), Frederick Schauer (University of Virginia), Mark Textor (King's College London), Achille Varzi (Columbia University), Wojciech Żełaniec (University of Gdańsk)

Argumenta 9, 1 (2023)
Target Article

Non-Persistent Truths

Andrea Bonomi

The Journal of the Italian Society for Analytic Philosophy

Non-Persistent Truths

Andrea Bonomi

University of Milan

Abstract

I start from Evans' criticism of temporalism, based on the claim that it does not "provide for the stable evaluation of utterances". I try to show that, with suitable qualifications, assuming the possibility of evaluations yielding different truth-values at different times is not an "eccentric" move (as suggested by Evans). I briefly consider Prior's metaphysical arguments in favour of the asymmetry between past and future and I suggest that, independently of these arguments, there are linguistic reasons in support of such an assumption. In particular, there are some future-oriented statements which (unlike past-oriented statements) are conceived of by speakers as intrinsically revisable and which require a non-monotonic characterization of the changing backgrounds of information selected by the time flow. As shown by some peculiar uses of phase adverbs like "still" and "no longer", variability in terms of truth-value assignation is a distinctive feature of this kind of statement. But another kind of variability of truth-value assignation is detectable in the case of present or past-oriented statements: in general, by refining the notion of context, it is possible to individuate different types of propositional contents, depending on which contextual parameters are abstracted over in order to account for different needs in communicative exchanges. Thus, in the final section of the paper, a more articulated notion of context allows for a richer (preliminary) description of the propositional contents that can be associated to utterances by abstracting over the relevant parameters.

Keywords: Radical temporalism, Asymmetry between past and future, Future-oriented statements, Multiple-choice paradox, Monotonicity.

1. An Eccentric Proposal?

In his criticism of Prior's tense logic, Evans (1985: 347) defines radical temporalism as a semantic theory according to which "the evaluation of particular utterances must change as the world changes". More exactly, he associates this form of temporalism with the following characterization:

(RT) □ *S* □ *u* □ *t* [Of(*S*,*u*) □ [Correct-at-*t*(*u*) □ TRUE$_t$(*S*)]]

where *S* is a variable for sentences, *u* for utterances and *t* for times. According to Evans, the problem, with such a characterization, is that it does not "provide for the stable evaluation of utterances as correct or incorrect": while all the

Argumenta 9, 1 (2023): 119-156
ISSN 2465-2334

© 2023 Andrea Bonomi
DOI 10.14275/2465-2334/202317.bon

utterances of *S* express the *same* proposition, the evaluation of an utterance as such is not fixed once and for all, because the proposition it expresses can have *different* truth-values at different times.

This kind of temporalism "is such a *strange position* that it is difficult to believe that anyone has ever held it". Indeed, according to Evans' reconstruction, what is not acceptable in (RT) is the fact that the evaluation of an utterance as correct or incorrect does not depend upon when the utterance is made, but may depend upon the evaluation time *t*, *whatever t* may be. This independence of the evaluation time with respect to the circumstance in which the utterance occurs would be the original sin of temporalism, because for the advocates of tense logic "to know what assertion is being made by an utterance all you need to know is *which* tensed sentence was uttered; you do not need further information to tie the tensed sentence down to a particular time [...]. It would follow that that such an 'assertion' would not admit of a *stable* evaluation as correct or incorrect" (Evans 1985: 349). In this passage, Evans endorses a stability principle which can be generically expressed as follows:

(SP) Let u be an utterance of a sentence *S* and t_u the utterance time:[1]

(i) *u* must be evaluated as correct or incorrect at t_u;

(ii) if *u* is evaluated as correct (incorrect) at t_u, then *u* must be evaluated as correct (incorrect) at any moment $t \neq t_u$.

In what follows, I will try to show that, with suitable qualifications, there are linguistic data showing that the stability principle (SP) is not always applicable and that a flexible notion of propositional content can help to account for the cases in which it fails.

2. Stability Forever

One way to get rid of the original sin described by Evans and to preserve the spirit (if not the letter) of the stability principle without resorting to eternal propositions is to assume that the correctness of an utterance *u*, in Evans's sense, depends on the truth-value that its content receives with respect to a *privileged* time of evaluation. And since any utterance *u* takes place at the utterance time t_u, thenatural solution is to say that t_u itself is the time span to which the evaluation of *u* as correct or incorrect must be anchored *once and for all*. Such a strategy would allow us to preserve the idea that a proposition (the content expressed by an utterance in the given context) can have different truth-values at different times, while the reference to a privileged time (t_u itself), *and to the world in which u occurs*, ensures stability in evaluating an utterance as correct or incorrect (or simply true or false).[2] As a matter of fact, in order to evaluate an utterance *u*, at t_u, of Geach's example (discussed by Prior and Evans)

(1) Socrates is sitting

what you have to do is simply to check whether the tensed proposition that Socrates is sitting is true at the utterance time t_u and in the utterance world w_u. If he is, the utterance is correct and will remain correct at any time $t > t_u$.

[1] The implicit assumption, here, is that *S* is no *deviant* sentence, in any plausible sense of the term.

[2] See footnote 25 for a justification of this way of speaking.

In principle, nothing changes if we consider utterances of sentences such as

(2) Socrates was sitting

or

(3) Socrates will be sitting.

As before, the correctness of these utterances must be evaluated with respect to t_u itself. The only difference is that *other* times, besides t_u, are involved: a time earlier than t_u, in the case of (2), and atime later than t_u, in the case of (3). So, an utterance of (2) is correct if Socrates is sitting at some time earlier than t_u, while an utterance of (3) is correct if Socrates is sitting at some time later than t_u. Far from being a problem, the fact that propositions have different truth-values at different times allows for a non "eccentric" way to deal with time and tense. This is possible because on such an approach the correctness of an utterance is evaluated, once and for all, with respect to the utterance time itself. Thanks to this anchoring effect, the utterance seems to admit of a *stable* evaluation as correct or incorrect, because, independently of the time flow, the evaluation time for the utterance remains fixed at the utterance time itself. Truth (or correctness), for an utterance, coincides with *truth in context*: this is the way in which "eccentricity" is avoided in Kaplan's semantics for tensed sentences.

Assuming for the sake of simplicity that the context c, for an utterance u, is represented by the time and the world at which u takes place (that is, $c = <t_u,w_u>$), the definition of correctness (or simply truth) for an utterance u of a sentence S can now be expressed as follows (where $[\![S]\!]^c$ is the proposition or intension expressed by S in context c, i.e. a function from pairs of times and worlds to truth-values):

(MT) $\Box S \Box u$ [Of(S,u) $\Box$ [Correct(u) $\Box$ $[\![S]\!]^c(<t_u,w_u>) = 1$]].

On this analysis, the evaluation of a given utterance in context c as correct or incorrect does *not* change in function of the time flow, even though the proposition it expresses in context c may havedifferent truth-values at different times. Stability is ensured since such an evaluation is anchored once and for all to the utterance time itself.[3]

The rest of the paper is organized as follows. I will briefly consider Prior's metaphysical arguments in favour of the asymmetry between past and future. I will then try to show that, *independently* of these arguments, there are *linguistic* reasons in support of such an assumption. *Pace* Lewis, the existence of open alternatives toward the future, but not toward the past, is not simply motivated by epistemic factors (our ignorance about future events), but is seen by speakers as

[3] Actually, this kind of solution à la Kaplan in order to preserve the Stability Principle is not accepted by Evans. As pointed out by Kölbel (2009), "Evans believes that the semantic values assigned by a semantic theory to sentences incontext should *immediately* and *as part of the semantic theory* yield evaluations of utterances as correct or incorrect". In particular, according to Kölbel, Evans rejects the following Kaplanian "bridge principle":

> An utterance of a sentence is *true* just if the content (intension) expressed by the sentence in the context of theutterance assigns the value true to the circumstance of evaluation *of* the context.

I will not go deeper into such issues, concerning the adequacy of Kaplan's approach as a way to preserve the Stability Principle, for the main goal of the present paper is to show that, *at least for a particular class of utterances, there is noreason to assume that principle*.

a characterizing feature of the way temporal determinations are semantically processed. This is why there are future-oriented statements which (unlike past-oriented statements) are conceived of as *intrinsically revisable* and which require a non-monotonic characterization of the alternative backgrounds of information selected by the time flow. As shown by some peculiar uses of phase adverbs like "still" and "no longer", variability in terms of truth-value assignation is a distinctive feature of some typical future-oriented statements and justifies the idea of an *evolving* context of utterance which inspires the semantics presented here. Depending on which contextual parameters are abstracted over, *different kinds of propositional contents can be individuated in order to account for the variety of conversational situations in which we refer to what is expressed by an utterance*.

3. The Utterance World(s)

It should be noticed that (MT) can work only if w_u contains *all* the necessary information with respect to *whatever* time may be involved by the tense in *S*. If, for instance, *S* is a future-tensed sentence like "It will be the case that ϕ" we have:

(4) "It will be the case that ϕ" is true at $<t_u,w_u>$ iff there is a time t such that $t > t_u$ and ϕ is true at $<t,w_u>$.

The point is that the temporal transition from t_u to t, in (4), has no effect on the choice of the relevant world, for just one single world (that is *the* utterance world w_u), with a single past and a single future, is associated to t_u. To evaluate the statement expressed by the utterance at issue, just look at what happens at some time in *this* world, exactly as you refer to some place in *this* world when a spatial location is involved.

This is exactly what is questioned by indeterminists like Prior. If the future, unlike the past, is open, evaluating an utterance of a future-tensed sentence at t_u involves a plurality of worlds or courses of events: those worlds that are all alike with respect to past and present events, while differing from each other with respect to the future (that is the worlds that are *metaphysically*[4] possible at theutterance time t_u, considering the events occurring at t_u and before t_u).

Prior's idea of the asymmetry between past and future can be illustrated by his reflection on what I called the multiple-choice paradox:[5]

> (MCP) Suppose *A* and *B* are being pushed towards the edge of a cliff, and there will be no stopping this process until there is only room for one of them. Then we may be able to say truly that it will definitely be the case that *A* or *B* will fall over, even though we cannot say truly that *A* will definitely fall or that *B* will definitely fall over (Prior 1957: 85).

Independently of the plausibility of this kind of example (a point on which I will return when discussing the role of the background of information in evaluating future-oriented statements), it isinstructive to follow Prior's argument.

[4] In the sense of Condoravdi 2001.

[5] I use this term because Prior's example is a future-tensed version of the "multiple-choice paradox" discussed in Bonomi 1997 (181-84) with respect to the progressive. Unfortunately, at that time I was convinced that this kind of argument should not apply to the future tense since I was not considering its modal import.

The problem, here, concerns *contingent* future events (such as being pushed towards the edge of a cliff and falling over), and Prior suggests to consider the *present* state of affairs as an appropriate criterion to distinguish, among the future-oriented statements, those that are *definitely* true (at the utterance time) from those that are not. As we have just seen, in his example this point is illustrated by the statement:

(5) *A* or *B* will fall over

which, according to Prior, turns out to be definitely true in the circumstances described above, whilst *neither*

(6a) *A* will fall over

nor

(6b) *B* will fall over

is definitely true in those circumstances (this is the apparent paradox).

In other words, in the above passage Prior's assumption is that the evaluation of future-oriented statements as definitely true or false depends on *present* facts or circumstances.[6] A statement like "It will be the case that ϕ" is true, at time t, if the truth, in the future, of ϕ is already *settled* at t.

One way to flesh out this notion of settledness is proposed by Thomason (1970): a proposition ϕ is settled, at time t, if ϕ is true in *every* course of events which is metaphysically (or historically, as hesays) possible at t. Thus, in particular, "It will be the case that ϕ" is settled at t if in each of those courses of events there is a time $t' > t$ such that ϕ is true at t'. Let us call *settledness* condition such a requirement.

It is also clear, from Prior's example, that settledness is a property of statements that *depends on time* in this sense: what is not settled at time t can *become* settled at a later time t' in view of new facts. (In the original example: at the beginning of the process, that *A* or *B* will fall over is not settled, but it *becomes* settled at some point in the process.) This point is made explicit, in connection with the so-called Peircean approach, by Prior (1967: 129): "'Will'" here means 'will definitely': 'It will be that *p*' is not true *until* it is in some sense settled that it will be the case, and 'It will be that not *p*' is not true *until* it is in some sense settled that not-*p* will be the case" (Italics mine).

The problem, at this point, is to know what makes the truth of a proposition *settled*. We have just seen that, on Prior's analysis, settledness rests on a *metaphysical* basis. Due to indeterminism, any moment t is associated to a multiplicity of future courses of events that are compatible with the events occurring at t or before t: settledness, for a proposition p, at a given moment t, means truth at all historical alternatives. The idea is that the future occurrence of the relevant event is, as he says, *unpreventable* at t.

One might challenge, of course, the plausibility of this analysis with respect to the semantics of future tensed statements in natural language, for the obvious question is: what makes a future event now unpreventable when we speak, for

[6] Øhrstrøm and Hasle (1995: 265) associate Prior's point of view to the following principle (where F is a "metric" future operator):

(P) The proposition $F(n)p$ is true now if and only if there exist now facts which make it true (i.e., which will make it true in due course).

example, of planned[7] events like a conference, a travel, an appointment or, simply, my next breakfast? If settledness is defined in terms of the metaphysical notion of "being unpreventable" in Prior's sense, then it can hardly represent a plausible necessary condition for the truth of future-oriented assertions, at least in a speaker's intuitions. What is missing here is the role that a background of information plays in determining what is settled at a given time.

Thus, after discussing some linguistic data, in the next sections a more flexible notion of settledness will be adopted in order to account for the role of the background of information in fixing the appropriate truth-conditions.

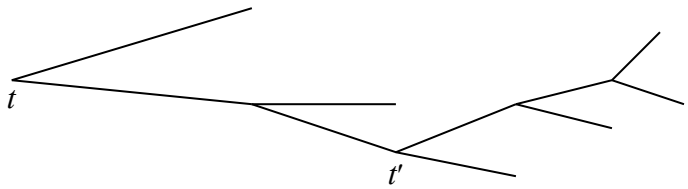
4. Monotonicity

As we have just seen, on Prior's analysis settledness depends on time, for the truth of a statement may be unsettled at time t, but settled at a time $t' > t$. The reverse is not possible, of course: the truth of a statement cannot be settled at t but unsettled at t', if $t' > t$.

In Thomason's formalization, such an approach is still conservative enough to meet the following requirement of stability: *if* the statement expressed by an utterance of sentence S is settled as true(false) at any time t, then it is settled as true (false) at every time $t' > t$. Let us see why.

As shown in Fig. 1, in the branching time (BT) framework associated to this analysis of tensed statements, the past moments, but *not* the future ones, are linearly ordered: given any moment t, there is only one course of events stemming from t towards the past, whilst there is a plurality of courses of events stemming from t towards the future.

Fig. 1



This is so because when you proceed from t toward the future, i.e., when you pass from t to a moment $t' > t$, new information gets available: which means that the metaphysical alternatives decrease (the branches stemming from t' are fewer than those stemming from t). In other terms, a BT model à la Thomason is monotonic in this sense:

(Mon) $t < t' \rightarrow H_{t'} \subseteq H_t$

where, for any moment x, H_x is the set of courses of events passing through x, that is the set of courses of events that are metaphysically possible at x.

An immediate consequence of (Mon) is that in such a framework stability of evaluation is respected in the following (weak) sense:

(WSP) (i) An utterance of a sentence S may fail to be evaluated as correct or incorrect (or simply true or false, as specified above) at the utterance time t_u or later.

[7] Such situations are extensively analyzed in Copley 2009.

(ii) But, once it has been evaluated as correct (incorrect) at a given moment t, it must be evaluated as correct (incorrect) at any moment $t' > t$.

This characteristic is inherited by the semantic system adopted by MacFarlane (2003, 2008), where the only possible transition is from neither true nor false to true (or false), but not from true to false or from false to true. (Actually, as far as I can judge, this kind of semantics is not designed to provide a unified treatment of the multiple interpretations that the future tense has in a natural language like English. The epistemic reading is just an example.) As in Thomason's approach, settledness, for future-oriented statements, is defined in terms of what happens in all the historical alternatives that are live options at the time of evaluation (or assessment). Once more, thanks to the monotonicity of the model, stability of evaluation is not questioned (starting from the moment at which an evaluation is possible).

5. Non-Persistent Truths: What We Know About the Future

Let us pause for a while. We have seen that, *under the assumption of the stability of evaluation*, Evans' criticism raises a problem for temporalism, according to which the content expressed by an utterance is a tensed proposition, in the sense that it is temporally neutral. We have also seen that a possible way out is to anchor the evaluation of this propositional content to a particular world (with a single past and a single future) and a particular time: the world and the time at which that utterance occurs. But such a solution, based on Kaplan's characterization of *truth in context*, is hardly compatible with indeterminism, i.e., a metaphysical orientation which has often represented one of the main theoretical motivations for temporalism and which associates an utterance event with a plurality of worlds (as far as the future is concerned). So, a natural alternative, at this point, is to accept the stability principle in a revised form (as stated in (WSP)), which is compatible with the fact that the evaluation of an utterance may be unsettled *until* the relevant conditions are fulfilled. Starting from this point, thanks to monotonicity, the evaluation of that utterance as correct or incorrect is stable, as desired.

This solution (which in Thomason's formalization is essentially based on a supervaluational approach) is an attractive way to cope with the issues raised by the adoption of an indeterminist point of view and to preserve (a revised formulation of) the stability principle, that is (WSP). Yet, *independently of our attitude toward indeterminism*, there is a preliminary question which should beaddressed if we are concerned with the semantics of the temporal markers in natural languages (of the future tense, in particular).

Are we justified in assuming that the evaluation of an utterance is stable (even in the weak sensestated in (WSP))? Does such an assumption conform to the intuitions (if any) of the speakers?

As a first step, consider the following example, inspired by a true story. Sandro (a good friend of mine) asks me whether it is *true* that I will leave tomorrow morning with the 6.45 train. My answer is that it is *true* (after all, I've already bought the ticket, made a reservation, packed my stuff, and so on). So, since he *knows* that I'm leaving with the 6.45 train, and since he is a generous man,

Sandro promises to take me to the station. Unfortunately, when he sets the alarm-clock, he makes a mistake. Conclusion: I miss the train. My comment is

(7) You knew that I would leave with the 6.45 train (you had to be more careful...).

The problem, in this case, is that, intuitively speaking, (8) is true at time t if there is a time t' such that $t' > t$ and the following statement is true at t':

(8) Sandro knows that I will leave with the 6.45 train.

On the other hand, it is an uncontroversial assumption that "know" is a factive verb which entails the truth of the propositional complement. So, what Sandro knows at t' cannot be *false*... But how is this possible, considering the fact that I did not leave with the 6.45 train?

To answer this question, take the following sequence of sentences:

(9) Leo knew that Lea would leave with the night train.

(10) So, he ran to the station and convinced her to leave with the morning train.

(11) Theo knows that Lea didn't leave with the night train.

The subordinate sentence in (9) is a further illustration of the future in the past (which in languages such as French or Italian can be expressed by an imperfective form or by a past conditional).[8] As before, a necessary condition for the correctness (or simple truth) of an utterance of (9) is that there is a moment t such that t is earlier than the utterance time and it is true at t that Leo knows that Lea will leave with the night train. (Let us assume, for instance, that she has already bought the ticket for this train, that she is on the right platform, etc.) Once more, this seems to be a very natural use of the verb "know" and, under the assumption that "know" is a factive verb, we must conclude that, if at time t Leo utters the sentence:

(12) Lea will leave with the night train

the statement made by this utterance must be evaluated as true, *at t itself*.
On the other hand, because of the factivity of "know", (11) entails that Lea did not leave with the night train: which seems to be in contrast with the correctness of Leo's utterance of (12). So, intuitively speaking, the same utterance must be evaluated as correct (to use Evans' terminology) at the utterance time t, but incorrect at the present moment: which is incompatible with the stability principle for utterances (even in its weaker version, based on monotonicity).

A possible objection to this kind of argument is that we cannot truthfully say that Leo knows, at t,that Lea will leave with the night train if Lea does not really leave with that train. For the same reason, the statement made by an utterance of (12), occurring at t, cannot be evaluated as true, at t itself, if the relevant event does not take place at the intended time. Thus, according to this

[8] A similar example, taken form a French magazine, is the following:

(K) DSK savait qu'il quittait les Etats Unis [DSK knew that he would quit the United States].

In this case, the future in the past is expressed by an imperfective form ("quittait"). Once more, the problem is the apparent contrast between the truth of (K) (which is genuinely asserted by the speaker) and the fact that the speakerherself is perfectly aware that Strauss-Kahn did *not* quit the United States, for he was arrested before leaving.

objection, (8) and (9) instantiate an improper use of the verb "to know", and the argument at issue should be rejected, while the stability principle can be preserved.

The natural answer to this objection is that it does *not* mirror the real behaviour of the speakers (and the corresponding intuitions) and the way future tensed sentences are used and evaluated (as true or false) in the appropriate circumstances.[9]

As a further illustration of this point, imagine the following scenario.

(i) On June 27 the Republican National Convention nominates Sarah Palin the official candidate for the 2012 Presidential Election.

(ii) On July 27 Sarah Palin is forced to give up because of her last hunting fiasco (she shot 285 times at a wandering caribou and missed).

(iii) On October 27, at the end of a new Republican Convention, Michael Moore is nominated the official candidate(and wins the Presidential Election).

Now consider the following sentences:

(13a) The person who will run for President in the 2012 Election is a woman (uttered on June 28)

(13b) The person who will run for President in the 2012 Election is no longer a woman (uttered on October 28).

From an intuitive point of view, (13a) would be judged as simply true, at the utterance moment u, by any competent speaker. This is so because, at u, the definite description "the person who will run for President" refers to Sarah Palin, not to Michael Moore. The obvious idea is that in such cases truth and reference do not depend upon the way the world will actually be, but upon the current (appropriate) information, for instance, about the relevant nominations.

The point is that this kind of information can change over time: this is why an utterance of (13b) does not mean, of course, that the candidate has changed sex (as predicted by the usual interpretation of "no longer"), but that something that was true in the past is no longer true at the utterance moment.

As for definite descriptions in particular, there is a clear asymmetry between past and future, for the reference of a future-oriented definite description can change over time, as shown by the fact that by uttering (13a) on June 28 we would make a true statement, whilst by uttering it on October 28 we would make a false one. On the contrary, the only natural interpretation of a statement like (13c) is that this statement entails a change of sex, not a change of truth value:

(13c) The person who ran for President in the 2008 Election is no longer a woman.

This contrast between past and future as concerns definite descriptions can be expressed by the following generalizations.[10]

[9] If Theo asks me "Is it true that Lea will leave with the train night?" and I reply "Ask Leo, he knows the truth" what I mean is not that he has improbable divinatory capacities and that he can read into the future, but, more plausibly, that he is provided with the right information about a planned course of events.

[10] The obvious assumption, here, is that the referent of the definite description does not depend on the presence of indexical expressions, for in such cases a past-oriented definite

(RefVar) It may happen that the referent of a future-oriented definite description (like "The person who will run for President in the 2012 Election") turns out to be the individual x at a given time t and the individual y ($y \neq x$) at a time $t' > t$.

(RefStab) If, at moment t, x is the referent of a past-oriented definite description (like "The person who ran for President in the 2008 Election"), then x is the referent of that description at any moment t' such that $t' > t$.

Notice that, *independently* of our philosophical assumptions about indeterminism, this contrast between an open future and a closed past as concerns truth and reference seems to mirror the way the future is conceptualized by the speakers when they use a sentence like (12) or (13a). It is the reference to a background of information about plans, motivated intentions, programs, etc., that explains why an utterance of (12) made at moment t can be evaluated as correct (true) at t itself, whilst it can be evaluated as incorrect (false) at $t' > t$, in view of new available information. Since the future, unlike the past, is (seen as) open, what is settled as true at t may not be settled as true at a time t' later than t, as shown by the fact that a sequence like (9)-(11) makes perfectly sense.

6. The Future in the Past

A crucial assumption, in the above argument, is that a statement like (9) which illustrates the so-called future in the past—is characterized by two important features: (i) the past tense takes us back to a past a moment t (that is $t > u$, where u is the utterance time); (ii) the relevant set of alternative futures is determined against a background of information *which holds at t itself*, not at the utterance time u. That is why, in the given scenario, statements like (9) and (11) are perfectly consistent.

This backward shift of the point of view involved by the future-in-the-past phenomenon is independently observed in other situations.

As an illustration, consider the following Italian sentences (uttered at a given moment u):

(14) Leo potrebbe (present conditional) partire domani mattina o domani sera (visto che ha fatto entrambe le prenotazioni). [Leo might leave tomorrow morning or tomorrow night (since he made both reservations)].

(15) Ma partirà domani sera. (Così incontrerà Lea a pranzo.) [But he will leave tomorrow night. (So, he will meet Lea for lunch.)]

According to a natural interpretation of (14), what the speaker means in this context is that, at the utterance time, there are aspects of reality, i.e., facts, which in principle make two alternative events possible. Thus, making a prediction on *contingent* issues (as in (15) a prediction based for instance on a given planning, a program, a reliable intention, etc.—is perfectly consistent with the awareness that a different course of events (with respect to that prediction) cannot be ruled out, as stated in (14). To put it in a slightly different way, the speaker

description can have different referents at different times. (Let us consider a definite description like "The person who bought me a drink yesterday night" which can designate individual a at moment m and individual b at moment m'). Crucially, the contrast between (13b) and (13c) concerns definite descriptions whose referents are fixed by dates.

seems to refer here to two distinct criteria to determine the intended universe of *possibilia*: (i) in (14) what is relevant is the universe of possible courses of events that are compatible with the way the world is at the present moment (the *metaphysical* possibilities, in the terminology adopted here, which are still open); (ii) in (15) this universe is restricted to the courses of events that are compatible with some extra assumptions about a planned course of events (so that only a part of those metaphysical alternatives are preserved).

Notice that the existence of the alternative options referred to by uttering (14) is seen by the speaker as *independent* of her epistemic preferences, according to which (witness (15)) only one option is to be selected.

But take these other sentences (uttered at *u*):

(16) Leo potrebbe (present conditional) essere partito ieri mattina o ieri sera [Leo might have left yesterday morning or yesterday night]

(17) Ma è partito ieri sera. (Così ha incontrato Lea a pranzo.) [But he left yesterday night. (So he met Lea at lunch.)]

Why does this sequence sound odd in Italian?[11]

Assuming an asymmetry between past and future in the way temporal information is encoded by thespeakers, here is a possible explanation of the contrast between the acceptability of (14)-(15) and the absurdity of (16)-(17).

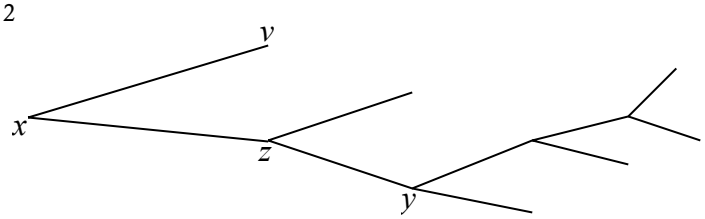
Given the present tense of the modal in (16), the reference time coincides with the utterance moment *u*. Under the hypothesis that, according to the speaker's intuition, what happened in the past (*unlike what will happen in the future*) is a *settled* issue, only one of the two alternatives mentioned in (16) is compatible with the current state of the world: in terms of metaphysical possibilities (Condoravdi 2001), only one option is open. Thus, the only plausible reading of the modal in (16) is the *epistemic* one: whether Leo left yesterday morning or yesterday night is a settled issue at the present moment, but I am unable to say what really happened. This is why, for all I know, *two* options are open. The problem is that this epistemic reading of (16) is not compatible with the statement made by (17), which presents one of the two options as definitely true. Hence the absurdity of the sequence.

To see this, consider Fig. 2. Suppose that the utterance time is located at *y* and that *z* represents a state of the world in which Leo leaves in the evening, whilst *v* represents the alternative state of theworld, in which Leo leaves in the morning. The past, unlike the future, is represented by a *single* path starting from

[11] Interestingly enough, the English sentence "Leo might have left yesterday morning or yesterday night", which is the natural translation of (16), is perfectly acceptable in this context, where it is followed by the sentence "But he left yesterday night". This is so, because "might" is compatible with a backward shifting of the perspective point from which *future* possibilities are considered. (See, on this point, Mondadori 1978 and Condoravdi 2001: in particular, her analysis of the ambiguity of a statement like "He might have won the game".) Thus, two metaphysical possibilities can bepresented as live options. But in Italian the *present* tense of "potere" rules out such a shifting, since the perspective point can only be located at the utterance time. The only possible interpretation of "potrebbe", in a sentence like (16), is the epistemic reading, but this reading is not compatible with (17), as shown in the text. That is why (16)-(17) sounds odd in Italian. (Indeed, such a sequence can be used to illustrate a sort of Moore's paradox: admitting that for what I know I cannot rule out the hypothesis that Leo left yesterday night is not consistent with the assertion that he left yesterday morning.)

y and this path includes only one of these alternatives, that is *z*. In other terms, *v* is no longer available as a metaphysical option from a perspective point located at *y* and can only be considered as an epistemic option. And since the epistemic reading of the modal in (16) is inconsistent with statement (17), there is no plausible interpretation of the sequence.

Fig. 2



Notice that, if in (16) the present tense of the modal verb is replaced by the past tense, the resulting combination is perfectly acceptable:

(18) Leo avrebbe potuto (past conditional) partire ieri mattina o ieri sera. [Leo might have left yesterday morning or yesterday night.]

(19) Ma è partito ieri sera. [But he left yesterday night.]

In this case, thanks to the *time shift* determined by the past tense of the modal verb, the perspective point is located at a moment which is in the past of the utterance time *y*, namely *x*, and at that time it *was* still possible that Leo would leave in the morning, even if such a possibility has not been actualized in the end. More exactly, in this case *v*, as a live metaphysical option,[12] is "accessible" from *x*, the time made relevant once the perspective point has been shifted: as a consequence, the modal in (18) is not forced to express an epistemic possibility (which would be incompatible with (19)), and the oddity disappears.

Conclusion: from a perspective point located at the utterance time *u*, open alternatives ("metaphysical" possibilities, to use Condoravdi's terminology) are available in the case of the future (as shown by the acceptability of (14)-(15)), but not in the case of the past (as shown by the oddness of (16)-(17)). In this case only epistemic alternatives are admitted. In order to make metaphysical possibilities available for the past, the perspective point is to be shifted to some moment in the past, so that the *future* of that moment is involved (witness the acceptability of (18)-(19)). Thus, if *the speaker's intuitions* are taken into account, there seems to be a difference betweenthe kinds of possibilities which can be associated, respectively, to the future and to the past: a plurality of metaphysical possibilities are admitted in the former case, but just a single metaphysicalpossibility is admitted in the latter case. As shown by (18)-(19). assuming a plurality of future alternatives, with respect to a given point in time, *is independent of an alleged state of ignorance*: after all (witness (19)), the speaker is provided with the correct information about the actual course of events.

This idea of a branching future and a linear past is a kind of asymmetry which does not depend on philosophical assumptions about indeterminism (so that we can stay neutral on this point), but seems to rest on a distinction

[12] According to Abusch (forthcoming), in such cases we should speak of "circumstantial" possibilities (in Kratzer's sense) and not of "metaphysical" possibilities. I leave this issue open for what I want to stress here is the need for a *backward shift* of the perspective point, independently of the nature of the *possibilia* associated to it.

underlying the semantic processing of tensed statements, whatever wemay conjecture about the nature of time.

As for the issue raised by Evans, since the open alternatives that are *contextually* relevant to evaluating future-oriented statements are sensitive to the time flow, the evaluation of a future-oriented statement can change as the world (with the associated expectations) changes, as we will see in the next section.

7. No Longer True

According to the program of tomorrow's concert, Bill Evans will play in a duo with Jim Hall. Leo, who has heard some vague rumours, asks:

(20) What about the tomorrow concert? Is it true that Bill Evans is playing with Jim Hall?

Since Lea is well informed, she promptly answers:

(21a) Yes, it is *true*.

(21b) Tomorrow Bill Evans is playing with Jim Hall.

As we have already seen, there is no doubt that such an answer testifies a quite intuitive use of the predicate "true" as applied to future-oriented statements and that it would be unnatural to object that, if the event at issue does not take place in the end, such a predicate is misplaced here. Once more, using this predicate in relation with a background of *current* information concerning a *planned* sequence of events (in the sense analysed in Copley 2009) is a fact that seems to mirror the speaker's intuitions, *independently* of philosophical speculations about the future and the debate on indeterminism.

Indeed, suppose that tomorrow, before the concert, the program is modified because of some unexpected events. According to the new program, Bill Evans will play with his trio. So, at this point Leo (who has been informed by the organizers of the concert) can call Lea before the concert and say:[13]

(22) Bill Evans is no longer playing with Jim Hall.

This is a very peculiar use of the phase adverb "no longer". In a different, and more familiar, kind of context an utterance of (22) would presuppose the existence of a *past* time at which an *event*[14] of Bill Evans' playing with Jim Hall was going on and would assert that such an event is not going on at the present time. But, since *no past event* of Bill Evan's playing with Jim Hall is involved in the scenario described above, what does Lea's utterance of (22) presuppose here? And what does it assert?

Roughly speaking, the idea is that this utterance of (22) presupposes that a *planning* about a certain kind of event was in force at some point in the past, whilst it asserts that such a planning is not in force at the utterance time.

The point is that there is an interesting relationship between (21b) and (22). Indeed, (22) can be analysed as follows:

(i) *presupposition* (triggered by "no longer"): the proposition expressed by Lea's utterance of (21b) [i.e., the proposition that Bill Evans will play with

[13] As I recall below, this kind of example is discussed in Dummett 2004. See Del Prete 2010 for a similar discussion about the examples suggested by B. de Cornulier and O. Percus (p.c.).

[14] Or series of events, on a common reading.

Jim Hall tomorrow night] *was true until* some moment in the past; it was true, in particular, at the moment of Lea's utterance (in the light of the original program);

(ii) *assertion*: this proposition is *not* true at the present moment (considering the new program).

Intuitively, the reason why the statement made by Lea's utterance of (21b) is true at the utterance moment u but false at a moment $t > u$ (witness the truth of (22)) is that these two moments are associated to two different backgrounds of information (based, respectively, on the original program and the modified program). In other words, the adverb "no longer" signals a change of the truth value which is to be assigned to the statement made by the utterance at issue, *depending on the moment* at which this statement is evaluated. The idea is that what is asserted by an utterance of a given sentence can be evaluated not only at the utterance moment itself, but at different moments, in function of the time flow. And since a transition from truth to falsehood (and vice versa) is always possible in the case of future-oriented statements, there is no reason to stick to the stability principle (not only in its stronger version, but also in the weaker one, according to which the only admissible transition is from neither-true-nor-false to a definite truth-value).

As a matter of fact, the content expressed by an utterance of (22) might also be expressed by an utterance of:

(22') It is no longer true that Bill Evans will play with Jim Hall

where it is evident that what we are evaluating *now* is the statement made by uttering (21b) at some past moment. So, a non-trivial consequence of this short excursus through the no-longer clauses is that the statement we make by uttering a sentence like (21b) in a given context is susceptible of evaluation not only in that context, but in a plurality of *changing* contexts, and that, as concerns *future-oriented* statements, there are clear cases of *variable* truth-values:

(TruthVar) It may happen that the statement made, in an appropriate context,[15] by uttering a future-tensed sentence turns out to be true (false) at a given time t, but no longer true (false) at a time $t' > t$.

This is what happens with statement (21b), witness (22) (or (22')).

Significantly, nothing similar happens with past-tensed sentences, as stated by the followingprinciple:

(TruthStab) It cannot happen that the statement made, in an appropriate context, by uttering a past-tensed sentence turns out to be true at a given time t, but no longer true at a later time $t' > t$.

As an illustration, consider a statement about the last week's concert like:

(23) Bill Evans did no longer play with Jim Hall.

As you recall, the natural interpretation of the future-oriented statement (22) is that it was true, at a past time t, that Bill Evans will play tomorrow with Jim Hall, and that this is no longer true at the present moment. But what about (23)?

[15] The assumption, here, is that there are no gaps in the information which is contextually required and that all the contextual coordinates have been fixed. For example, in the case of (22), or (22′), it must be clear from the context thatwe are speaking of the tomorrow concert. This point will be made clear in Sect. 12.

Of course, there is no possible interpretation of this past-oriented statement according to which, in analogy with the above interpretation of (22), it was true, yesterday, that in the last week's concert Evans played with Jim Hall, and that this is no longer true at the present moment.[16] And this seems to be an important asymmetry between past-oriented statements and future-oriented statements.

8. Still True

The moral we can draw from the examples we have just discussed is that the stability principle makes sense for statements about the past, but not for statements about the future. As remarked by Dummett in *Truth and the Past*, this conclusion about future-tensed sentences *does not depend on philosophical premises*, but is motivated by observation: "Independently of metaphysics, we incontrovertibly have a use of future-tense statements under which they are rendered true or false by how things stand in the present. This is exemplified by a statement 'They were going to be married, but they are not going to *any longer*'" (2004, italics mine.).

The existence of situations in which the evaluation of a future-oriented statement depends on "how things stand in the present" and, as a consequence, yields different results at different times, can explain some typical uses of *still*-phrases, which are so to speak "symmetrical" with respect to *no-longer*-phrases, as shown by the following example:

(24) A: Bill Evans might play with his usual trio tomorrow night and not with Jim Hall. I've heard that some of the organizers wanted to change the program.

(24b): Yes, they discussed about a possible change, but, for practical reasons, the program has never been modified. *So, Bill Evans will still play with Jim Hall.*

In this scenario the statement made by an utterance of a sentence such as

(25) Bill Evans will play with Jim Hall

is true at the utterance moment u and confirmed as true at a further moment $t > u$ in the light of the most recent developments. On the other hand, this kind of confirmation, expressed by (24b), makes sense only if we assume that evaluating the content of an utterance of (25) can yield different resultsin function of the time flow, depending on the background of information which is made relevant by facts *and* assumptions about planned events.

Intuitively, the semantics which has often been associated with "still" is the following (Katz 2003; Krifka 2000):

(Still) (i) if uttered at time t, "still P" entails that P is true at t;
(ii) presupposes that P was true at some salient time t' before t;
(iii) and that P has been true at all the times in between t and t'

As for (24b) such truth-conditions entail that the statement that Bill Evans will play with Jim Hallis true at the present moment, and presuppose that it has always been true, in the above scenario, even though such a possibility could sound problematic at some point.

[16] Dummett (2004) discusses the absurd content expressed by uttering the sentence "She then married Edward in 1825, but did not now do so".

More in general, phase adverbs like "still", "no longer", etc., in this very peculiar use, can occur in asentence in order to signal the effects of a change of the background of information on the evaluation of a given propositional content: roughly speaking, one presupposes the existence of a given background *X*, and one specifies what happens (in terms of validation/invalidation) to that propositional content after a transition to the background *Y*.

This peculiarity can be intuitively explained as follows: on the familiar interpretation, one concentrates on the effects of a *temporal* transition (i.e., when passing from moment *t* to moment *t'*)in terms of the continuation/termination of a given *event* or *state*; on the interpretation under discussion, one concentrates on a change in the background of information to see its effects on the evaluation of a given statement.

This is a general phenomenon which does not concern only temporality. For example, take a situation in which we are considering the possible changes of a fictional character (e.g., Major Amberson) when passing from a particular background of information (Booth Tarkington's original story: *The Magnificent Ambersons*) to another one (Orson Welles's film with the same title). In this case the following statements:

(TS) Major Amberson is no longer an arrogant man
(WS) Major Amberson is still an arrogant man (but at the same time he has a very visible side that renders him considerably more sympathetic)[17]

are perfectly acceptable in order to mean that what is true (about this character and his arrogance) with respect to the background of information provided by the original story is no longer (still) true with respect to a *different background*, represented by Welles' film. This interpretation of "no longer", for example, is quite different from the (more familiar) interpretation according to which in the novel *itself* Major Amberson is arrogant until moment *t* and no longer arrogant after moment *t*.

Going back to tensed sentences, examples such as (22) and (24b) show that, unlike past-oriented statements, future-oriented statements are conceived of by speakers as *intrinsically revisable*, depending on the changes which may occur in the flow of information about the world. The idea is that, in such contexts, a no-longer-phrase signals a change of truth-value due to a modification of the relevant background of information, whilst a still-phrase signals a persistence of truth value. But both phrases make sense, in these scenarios, only under a *defeasibility assumption* concerning the relevant proposition (the proposition that Bill Evans will play with Jim Hall in the tomorrow concert).

Such a defeasibility assumption may even be part of the explicit content expressed by an utteranceof a future-oriented sentence, as shown by the following example:

(25)-Next year the Olympic Games will take place in China. But in an emergency, the Games will be cancelled.

Indeed, it is not difficult to imagine a scenario in which (26) would be perfectly acceptable. On the contrary, there is *no* plausible situation in which (27) would turn out to be consistent.

[17] Thanks to O. Percus for suggesting a modified version of this example.

(26) Last year the Olympic Games took place in China. But in an emergency, the Games were cancelled.

A reasonable explanation for the contrast between these two discourses is based on a particular kindof transition concerning the representation of the open possibilities. When she evaluates the first sentence in (25), the hearer is invited to consider a restricted set of open alternatives: those which are compatible with the relevant contextual assumptions (e.g., the CIO's decisions). However, when she comes to the second sentence in (25) and processes the phrase "in an emergency", she shifts to a different set of alternatives (those in which something unexpected has occurred) and evaluates the sentence "The Games will be cancelled" relative to this shifted domain. This *change* in the domain of available alternatives explains why the second sentence of (25) does not contradict the first one.

But in the case of the past-tensed discourse (26) such a mechanism of transition cannot apply, because there is only one course of events relevant to evaluating past-tensed sentences, no alternative is available. If it turns out that the Olympic Games took place in China in the unique past available, then the possibility that the Olympic Games did not take place is not an open option. That is why (26) turns out to be inconsistent.

It is worth noticing that, if in (26) the reference to the past is replaced by the reference to a *future* in the past, what we get is a perfectly acceptable statement:

(27) Last year the Olympic Games took place in China. But in an emergency, the Games would have been cancelled.

What changes when passing from (26) to (27)? The idea is that, by replacing a simple past with a past *conditional*, one makes the future-in-the-past interpretation available: one refers to a past moment t in the *future* of which *alternative* courses of events stemming from t are relevant. This is why a transition between different sets of open alternatives is possible here, as in the case of (25). One often suggests that there is no difference, in principle, between future-oriented and past-oriented statements as regards the way they are semantically processed. The illusory asymmetry which associates the future, but not the past, to a plurality of alternative options is to be explained interms of *epistemic ignorance*. Semantically speaking, this is the conclusion: there is just a single future exactly as there is just a single past. As D. Lewis warns us, "the trouble with branching exactly is that it conflicts with our ordinary presupposition that we have a single future. If two futures are equally mine, one with a sea fight tomorrow and one without, it is nonsense to wonder which way it will be—it will be both ways—and yet I do wonder […] Our future is the one that is part of the same world as ourselves" (Lewis 1986: 207-208.) Plurality of options, one suggests, is begat by ignorance: it is only because we cannot have epistemic access to this single future that wetreat it as "open" and that a multitude of possibilities is associated with it. So, there is no intrinsic difference, this is the conclusion, between the single past involved by past-oriented statements and the single future involved by future-tensed statements.

A moment's reflection is sufficient to show that the data we have discussed so far suggest a more articulated view. The case of the future in the past, illustrated by (27), is particularly interesting in this connection, because the existence of open alternatives, toward the future, at a past moment t, is not due to lack of

information (the speaker knows what happened), but is seen as a characterizing feature of *that* moment. Symmetrically, the lack of alternatives, toward the past, is seen as a characterizing feature of the *present* moment.

In other terms, the contrast between (26) and (27) seems to suggest that, if we stick to the way temporal information is encoded in a natural language such as Italian or English, the past, but not the future, of a given time *t*, is inherently associated to the idea of a single course of events stemming from *t*.[18] As in the case of the contrast, discussed above, between sequence (14)-(15) and sequence (16)-(17), the idea is that the indeterminacy of the past can be justified only in terms of an epistemic failure, whilst the indeterminacy of the future does *not* coincide with a simple lack of information and is hardly compatible with the alleged "presupposition that we have a single future".

9. Evaluating Utterances in a Changing World: A First Approximation

The linguistic evidence we have discussed so far seems to suggest the following conclusions:

(i) As shown by the way the predicate "true" is used by speakers in relation to some future-oriented statements, the statement made by an utterance of a sentence like (12), (21b) or (25) is evaluated as true, at the utterance moment *u*, by referring to a relevant background of information (let us call it *VIEW* for brevity), whatever course of events may be actualized in the end.

(ii) There is an asymmetry between past and future, in the sense that while a single course of events is referred to for the evaluation of a past-oriented statement, in the case of a future-oriented statement a plurality of alternative courses of events is made relevant: it is the set of courses of events that are compatible with *VIEW*. As shown by the future-in-the-past phenomena and by some uses of epistemic modals, this asymmetry is seen by speakers as a constitutive feature of their representation of time and not as a simple product of our ignorance about future events.

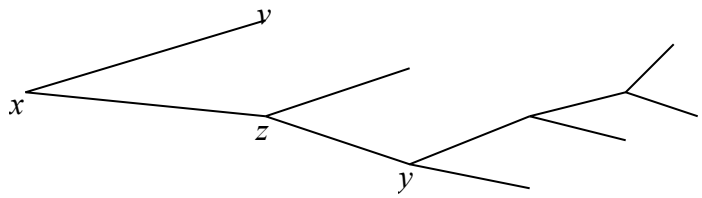
(iii) Some peculiar uses of phase adverbs like "no longer" and "still" show that the statement made by uttering a future-tensed sentence can be evaluated not only at the utterance time u, but at any moment later than u and that different evaluations are possible at different moments (because of the variability of *VIEW*). In other terms, this kind of statement is intrinsically defeasible, for the variability in truth value is not limited to the transition from an indefinite truth value to a definite one, but allows for the transition from truth to falsehood (and vice versa).

The problem, at this point, is how to flesh out such requirements in a suitable formal framework.

As a first approximation consider Fig. 2 once again:

[18] An important qualification is in order here. Insisting on this kind of asymmetry between past and future in the light of the data provided by examples like (26)-(27), (14)-(15) or (16)-(17), does not entail that such an asymmetry characterizes the structure of the time as such, but that it characterizes the way temporal information is processed by speakers in the production and interpretation of utterances.

Fig. 2



In this kind of representation, moments such as *x*, *y*, *z*, ... have a *double* role to play, according towhether we consider the tree on which they are located (*A*) on the vertical axis or (*B*) on the horizontal axis.

(A) A moment in the tree is a point which is alternative to other points in a logical space. (For example, in Fig. 2, *v* and *z* are alternative outcomes of the node *x* that precedes them.) An important characteristic of these points is that each of them can be uniquely associated to a plurality of histories. More exactly, for any moment *m*, let H_m be the set of histories passing through *m*: i.e., the histories that coincide up to *m* and diverge starting from that point. Thus, in what follows, when I intend to stress this aspect, I will refer to a moment *m* as a world or a world state (Prior) with a single past and a set of alternative futures (corresponding to the different histories in H_m).[19]

(B) But, of course, a moment *m* is also associated to temporal information and can be seen as a particular time, which precedes or follows other times. For example, in Fig. 2, *z* is earlier than *y* ($z < y$).

In this theoretical framework, if other contextual features are ignored, it is possible to consider a context as involving a pair of moments <*u*,*v*>, where *u* and *v* play distinct roles, because they are associated to the utterance time and to the utterance world (in the sense clarified in (A)), respectively. (An interesting illustration of this point is the pair <*u*,*u*>, where the same moment plays these two roles. Let us call it *the canonical context*.)

To grasp the intuition underlying such an approach, suppose that a sentence is uttered at moment *u*. Thus, the utterance time is fixed once and for all: it is *u* itself. But what about "the world" of the utterance? Surely, at moment *u*, *u* itself can be considered as a world in which the utterance event can be located, i.e., as the world of the context (with a single past and the alternative futures in H_u).This is the *canonical context* <*u*,*u*>. Yet, as time goes by, *other* worlds become available: for example, world *z* (or, alternatively, world *v*), because the utterance event at issue belongs to *this* world in the following sense.

An event *e* is said to belong to world *x* if *e* occurs at some point in the path up to and including *x*.[20]

An obvious principle of persistence can be stated in this connection:

[19] Formally speaking, I will identify the world (corresponding to) *m* as the particular subtree branching after *m* but linearly ordered up to *m*, i.e., as a cluster of temporally complete courses of events.

[20] See Bonomi and Del Prete 2008 for a more accurate representation of events in a BT framework, which is not within the scope of the present paper.

(PP) For any event e and for any moments x and y: if e belongs to world x and $x \leq y$, then e belongs to world y. (Intuitively speaking, in a changing world, a fact remains a fact.)

Thus, whilst the utterance time remains fixed, different worlds (in the sense relevant here, e.g., u itself, or v, or z, and so on) can in turn be considered "worlds of the utterance". Crucially, since principle (PP) guarantees that the utterance event (with the agent, the place, etc., of that event) belongs not only to u, but to any x such that $u \leq x$, referring to a standard definition of *proper* context is sufficient to state the following *Conservativity Principle* (CP).

Let a *proper context* for an utterance event e be a quadruple $<t, s, p, m>$ such that e belongs to world m (the world of e), s (the speaker of e) is located at p (the place of e) at the time t (the time of e) (see Kaplan 1977: 509). For any context c, let $c(w)$ be the world of c. In the light of these definitions, it is immediate to see that (PP) entails (CP).

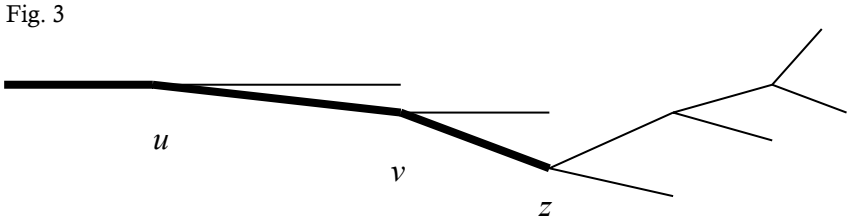
(CP): If e is an utterance event and c is a proper context for e, then c' is a proper context for e, too,where $c = c'$ except that $c(w) \leq c'(w)$.

In other terms, if in a proper context for an utterance event e the world of the context $c(w)$ is replaced by a "development" of $c(w)$, what we obtain is still a proper context for e. This fact will play an important role in the analysis which will be developed in the next sections and which is based on the idea that *a family of contexts should be associated to the utterance at issue*, dependingon which world is made relevant by the time flow.

To see this, for the sake of simplicity let us temporarily consider contexts as ordered pairs of type $<u,v>$, where u plays the role of the utterance time and v the role of the utterance world, respectively (do not forget that, in such an analysis, the same kind of entity can play two distinct roles, as shown by the canonical context $<u,u>$).

Fig. 3 can be helpful to illustrate this point. u, v, z are "worlds" to which the utterance event belongs and $<u,u>$, $<u,v>$ and $<u,z>$ are possible contexts for that utterance.

Fig. 3



Suppose that a sentence S is uttered at u, which means that one feature of the context is fixed once and for all: it is u itself. But u is also available in order to fix the second feature of the context: as a consequence, a first admissible context is represented by the pair $<u,u>$, *the canonical context*. Yet,in the light of principle (CP), *other* admissible contexts become available as time goes by, for instance $<u,v>$, or $<u,z>$ and so on.

This seems to be a very natural way of characterizing the notion of an *evolving* context of utterance,for it is quite intuitive to think that an utterance, like any other event, has effects which stretch along the time line(s). In particular, whenever an utterance entails a reference to a background of assumptions, it comes

as no surprise if different states of information are involved in function of the time flow.

To sum up, let us survey the main features characterizing this tentative analysis.

(i) The parameters composing a context are the usual ones (utterance time and utterance world, in the simplified and provisional version I have just sketched).

(ii) As suggested by the adoption of branching structures, a world is represented not by a single history, but by a moment x, seen as a cluster of histories (i. e. the histories passing through x).

(iii) The utterance time is uniquely fixed.

(iv) The utterance world is not uniquely fixed.

(v) Because of (iv), an utterance is associated not to a single context, but to a plurality of admissible contexts. More exactly, for any utterance event e and moment u such that u is the utterance time of e, the class of admissible contexts for e is the set of pairs $<u,x>$, where x can be replaced by any v such that $u \leq v$.

10. More Articulated Contexts

Crucially, on this tentative analysis, an utterance context is conceived of as a dynamic reality whichis sensitive to the time flow. When passing from x to y (where $y > x$) there is a contraction in the setof metaphysical possibilities that are still open.

Unfortunately, as we saw above, this kind of context variance is sufficient to account for a kind of evaluation according to which (the statement made by) an utterance may turn out to be neither true nor false at moment x and true (or false) at moment $y > x$, but is not sufficient to account for situations where we pass from truth to falsehood (and vice versa). But this is what may happen to the statement made by an utterance of a future-oriented sentence, which involves a background of information, as we saw when discussing the relevant examples.

To account for such situations, we have to associate, to any utterance u, not only the set of metaphysical possibilities open at u (corresponding to the histories in H_u), but also different backgrounds of information (what we called VIEW), in function of the time flow. In what follows the interaction between H_u and VIEW will be reconstructed by referring to a "system of spheres" which is a modification[21] of the one introduced by Lewis (1973) and which will cope with *non-monotonic changes of information*, as required by the possible transition from a definite truth value to its opposite. Here are the formal definitions.

Branching Time Models

A *BT model* is a structure $M = \langle T, \leq, D, F \rangle$, where[22]

[21] This version of Lewis's system is presented by Grove (1988) in a different theoretical framework. In what follows, I will consider systems of spheres centred on H_u, which is the universe of possibilities originally associated to the utterance time. Alternative choices are possible in other cases (for example to account for other modal operators).

[22] See Belnap et al. 2001 for the formal and philosophical aspects of this modelling.

(i) T is a non-empty set, the domain of *moments*, assimilated here not only to points in a temporal grid, but also to points (situations or *world-states* in Prior's sense) in a logicalspace. (See the discussion at the end of section 9.)

(ii) □□ is a partial order over T (i.e., it is a reflexive, transitive and anti-symmetrical relation over T); □ is forward branching but not backward branching (i.e., it is branching towards the future but not towards the past), as required by the following postulate:

□ m_0,m_1,m_2 [[m_0 □ m_2 □ m_1 □ m_2]□□ [m_0 □ m_1 □ m_1 □ m_0]];

(iii) D is a domain of individuals.

(iv) F is the interpretation function mapping predicates to their denotations relative to moments in T.

A *history* is a maximal □-chain on T. This mean that a set X□□ T is a history in T if X satisfies thefollowing conditions:

(i) □ m_0,m_1□ X [m_0 □ m_1□ m_1□ m_0]

(ii) □ Y□ T □ m_0,m_1□ Y [[m_0 □ m_1□ m_1□ m_0] □□ X□□ Y□ $X = Y$].

Intuitively speaking, histories are temporally complete linear paths, each of which can be seen as adeterministic course of events.
For any moment m□□T, H_m is the set of histories containing m.

Systems of Spheres

Let H be the set of all histories in a structure M = □ T, □, D, F□□ and, given a moment u, let H_u be the set of histories passing through u. A collection S of subsets of H is a *system of spheres* centred on H_u if it satisfies the followingconditions:

(i) S is totally ordered by □□

(ii) H□ S (as a consequence, H is the largest element in S).

(iii) H_u□ S and, for any B in S, H_u□ B (i.e., H_u is the □-minimum of S)

(iv) For any non-empty set of histories p there is a smallest sphere B′ such that $B' \cap p \neq \emptyset$. (This is the limit assumption discussed by Lewis.[23])

In virtue of (i)-(iv), a system of spheres S centred on H_u can be associated with a function □$_u$ which maps any non-empty set p of histories to a set of histories defined as follows:

□$_u(p) = E \cap p$, where E is the smallest sphere in S such that $E \cap p \neq \emptyset$.

Intuitive meaning. Suppose that u is the utterance moment. Thus, the smallest sphere H_u (on whichthe system is centred) is the set of histories that are metaphysically possible at u. In a sense, H_u "sets the scene" *by determining the horizon*

[23] This assumption, which is made here for the sake of simplicity, is rejected by Lewis (1973) as incompatible with his interpretation of the spheres. A system of spheres, in his original proposal, is centred on a single world @. As a consequence, if any given sphere B is chosen as the smallest sphere X such that $X \cap p \neq \emptyset$, it is possible to find worlds that are closer to @ than those in B: which contradicts this choice. As a partial justification for accepting the limit assumption one might argue that a system of spheres is centred here on a set of histories selected by a background of information and that worlds which are too finely individuated to be discernible with respect to this background might be treated as equally "close" to the centre of the system. (See Bonomi 2006.)

of possibilities with which different backgrounds of information can interact in order to select the relevant alternatives.

Given two histories h and h', if there is a sphere B such that B contains h but not h', we can say that h is "closer" to H_u than h', i.e., closer to the idea of what is metaphysically possible at u. For any non-empty set of histories X, $\Box_u(X)$ is the set of histories in X which are "maximally" close to H_u. As we will see in a moment, moving from the centre to the outer spheres, in combination with function $\Box_u$, will serve to account for the progressive availability of different backgrounds of information at different moments, starting from u. These new scenarios are determined by suitable *revisions* of the relevant information and are located at different levels of closeness to the original scenario. As desired, the structure is non-monotonic in the following sense: for any two moments x and y such that $x < y$, it may happen that $\Box_u(Y) \not\subseteq \Box_u(X)$ even if $H_y \subseteq H_x$, where X and Y are the informational backgrounds associated to x and y, respectively. (The contention is that, for any moment v,the background of information contextually selected for v is determined not only by the way the world is at v, but also by a relevant set of assumptions.)

Contexts

In order to focus on the core of the present proposal, I will ignore the features that are not relevant here by reducing a context c, for an utterance event e occurring at moment u, to the triple $c = \langle u, TT_c, VIEW_u \rangle$, where u is the *utterance moment*; TT_c is the time which is spoken about (a notion that will be discussed later on); $VIEW_u$ is the *reference time function*:[24] for any moment v such that $u \leq v$, $VIEW_u(v) = \langle p_v, S_u \rangle$, where p_v is the relevant background of information holding at v (or, more exactly, the set of histories in H that are compatible with such a background) and S_u is a system of spheres centred on H_u.

Thus, for any moment v such that $u \subseteq v$, $VIEW_u(v)$ can be associated to a particular set of histories, i.e., the set $\Box_u(p_v)$, where $\Box_u$ is the function associated to S_u described above. We will denote by "$VIEW_u(v)$" this set (that is, "$VIEW_u(v) = \Box_u(p_v)$, where $VIEW_u(v) = \langle p_v, S_u \rangle$. Intuitively speaking, $VIEW_u(v)$ is the set of histories which, *in the background of information holding at v*, come closest to idea of what is metaphysically possible at the utterance moment u. As we have justspecified, $VIEW_u(v)$ is determined not only by the way the world is at v, but also by such a background (which might include the reference to plans concerning future courses of events, for example). Because of this changing informational content, it may happen that $VIEW_u(t') \not\subseteq VIEW_u(t)$, even though $t < t'$ and $H_{t'} \subseteq H_t$. As we shall see in a moment, thanks to this lack of monotonicity concerning the set of *possibilia* associated to different moments, a future-oriented statement can be evaluated as true (false) at a moment t but false (true) at a moment $t' > t$.

Notice that, due to the presence of the reference time *function* $VIEW_u$, an utterance context c has an inherently *dynamic* character. Indeed, for any context

[24] To simplify, given the examples under discussion, in the present context the reference time function applies to moments coinciding with the utterance time u or following it. However, in an extended theoretical framework nothing prevents it from being associated to moments preceding u.

$c = <u, TT_c, VIEW_u>$, the utterance time is fixed once and for all, and is represented by u, which is the moment of the utterance event. But what about function $VIEW_u$, which selects the relevant background of information? Surely, the moment uitself can play the role of evaluation moment to which $VIEW_u$ applies. This means that $VIEW_u(u)$ is associated to a particular observation point. Yet, as time goes by, *other* moments become available as moments which feed function *VIEW*: for example, moment v, or, later on, moment z, and so on, so that, by suitable *revisions*, other backgrounds of information can become available in the same utterance context. As we saw above, the intuition, here, is that, once an utterance event e has taken place, *the effects of this event stretch far along the time line,* as represented in Fig. 3, where u, v, z, …, are ideally associated to different world states and to different backgrounds of information:[25] as a consequence, the new perspective points may involve possible courses of events that were previously ruled out. (While the metaphysical possibilities decrease when passing from time t to time $t' > t$, the universe of possibilities associated to $VIEW_u(t')$ is not necessarily included in the universe associated to $VIEW_u(t)$.)

11. Back to Non-Persistent Truths: The Utterance World as a World in Progress

After presenting the idea of an *evolving* context of utterance, in which different observation points can be referred to at different moments and the change of informational background is non-monotonic, we are in a position to fix the truth-conditions of a statement made by an utterance of a future-tensed sentence[26] and to show how the evaluation of such a statement may *not* obey the stability principle discussed at the outset. (For brevity, from now on I will speak

[25] What Kratzer writes about modalized sentences seems to apply to the treatment of future-oriented sentences proposed in this paper: "We might wonder why there should be a unique conversational background for a modalized sentence to express a proposition. This seems too strong. More often than not, conversational backgrounds for modal remain genuinely *underdetermined* and what speakers intend to convey is compatible with several choices of conversational backgrounds" (Kratzer 2012: 323; italics mine). In the case of the future tense, I suggest that we speak of a sort of *announced indeterminacy* as concerns the background of information, which is to be fixed by the context, in the sense that, as time goes by, *different* backgrounds can be associated to *different* moments in the *same* utterance context.

[26] In what follows the future tense is associated to a sentential operator along the lines of traditional Priorean treatments. This choice makes a comparison with those treatments easier. Actually, the main idea developed in the present paper (and based on a "dynamic" characterization of the utterance context) is compatible with (or even more attuned to) other choices, in particular with a referential treatment of tenses in the spirit of Partee (1973) and Heim (1994). It is in this referential framework that future-tensed sentences are dealt with in Bonomi 2010, where a full compositional semantics is based on a richer notion of *utterance context* (involving not only a coordinate for the perspective point associated to $VIEW_u$, but also a coordinate for the target time). Del Prete (2010) proposes a modelling in which the future *per se* has no quantificational force: "a bare future sentence is interpreted by default in such a way as to have the temporal variable instantiated on every accessible future. The default interpretation of a future sentence is thus a universal quantification over a domain of accessible futures". For the sake of simplicity, I ignore these possible refinements and maintain the Priorean approach.

of the truth of an utterance.[27] This should capture the idea of "correctness" that Evans discusses in connection with utterances.)

Let a context c be the triple $<u, TT_c, VIEW_u>$ as defined above, where, in particular, for any v such that $u \leq v$, $VIEW_u(v) = <p_v, S_u>$, so that $VIEW_u(v) = \square_u(p_v)$. Let F be a sentential operator and v a moment in T. The truth of an utterance of a future-tensed sentence "$F\square$", in the context c, relative to moment v (and assignation g) is defined as follows:

(TCF):

$[\![F\square]\!]^{c,g,v} = 1$ iff $u \leq v$ and $\square h'\square (VIEW_u(v)) \square v'\square h'(u < v' \square\square [\![\square]\!]^{c,g,v} = 1)$
$[\![F\square]\!]^{c,g,v} = 0$ iff $u \leq v$ and $\square h'\square (VIEW_u(v)) \square v'\square h'(u < v' \square\square [\![\square]\!]^{c,g,v} = 0)$
Otherwise, $[\![F\square]\!]^{c,g}$ is undefined.

Suppose that an utterance of "$F\square$" occurs at moment u, where u is a coordinate of context c. According to (TCF), this utterance is true at a moment v (coinciding with u or later than u) iff $\square$ is true at some moment v' later than u in all the histories which are compatible with the background ofassumptions holding at v and which are "maximally" close to H_u.

To see how the truth conditions in (TCF) allow for non-persistent truths, let us go back to example (21b):

(21b) Tomorrow Bill Evans is playing with Jim Hall.

As we saw when discussing this example, the statement made by an utterance of (21b) can be true if evaluated at the utterance moment u, in view of the original program for the concert, but false at a moment v, such that $u < v$ and v is later than the moment at which the program is modified (but earlier than the time at which the concert takes place). That is why

(22) Bill Evans is *no longer* playing with Jim Hall

or

(22') It is no longer true that Bill Evans will play with Jim Hall

can be truthfully uttered at v.

On the proposal under discussion, we would say that this is possible because there is a *change* of perspective when passing from moment u to moment v, and such a change is formally accounted for by the fact that function *VIEW* can associate different backgrounds of information to u and v, respectively. In other words, to account for the change of evaluation expressed by (22) or (22') we can simply say that the proposition expressed by (21b) *in the given utterance context c* turns out to be true at u, but false at v:

$[\![(21b)]\!]^{c,g,u} = 1$
$[\![(21b)]\!]^{c,g,v} = 0$

[27] Given an utterance of a sentence S in a context c (which includes the utterance moment u), it is possible to speak of the truth of that utterance (where Evans speaks of "correctness") *with respect to a moment v* in the following sense: the content (the proposition) expressed by that utterance in context c is true at v, that is $[\![S]\!]^{c,g}(v) = 1$. After all, such a definition of truth (correctness) for an utterance comes as no surprise with respect to a familiar kind of intensional semantics, where the truth of an utterance, in a context c, is relative to a circumstance of evaluation (world and time). What is new here is the fact that c does not associate a single background of information to that utterance but makes it dependent on the evaluation time (as required by some peculiar uses of phrase adverbs).

We can have different truth values because the intended proposition $[\![(21b)]\!]^{c,g,}$ is evaluated relativeto different moments (*u* and *v*, respectively), which in turn correspond to different backgrounds of information. As I have just recalled, in the formal framework under discussion this peculiarity is accounted for by associating the reference time with a *function*, which picks out different backgrounds depending on the time flow. More exactly, given an utterance context $c = <u, TT_c, VIEW_u>$, this task is achieved by its third coordinate, function $VIEW_u$, which represents the dynamic side of *c*, for it makes different moments available in order to evaluate the propositional content *with respect to that utterance context*. The point is that such a context determines not only the temporal location of the utterance event itself (which is fixed once and for all by the first coordinate), but also, thanks to function $VIEW_u$, the alternative moments or world states (with the associated backgrounds of information) which are relevant to the evaluation process. For the reasons discussed above (in connection with the conservativity principle), each of these moments isto be considered as a world of the utterance at issue or, if you prefer, a single world is involved here, but a changing one.

Specifically, the change of perspective justifying the contrast between (21b) and (22) is explained as follows

(25) $VIEWu(u) \neq VIEWu(v)$

where $VIEWu(u)$ is the set of histories compatible with the original program for the concert (whichis *still* valid at *u*), while this program is *no longer* valid at *v*, so that $VIEWu(v)$ selects the histories in which Bill Evans does not play with Jim Hall but with his trio.

Thus, we have detected an important source of contextual dependency, because the truth of an *utterance* is relative to the background of information selected by the reference time function $VIEW_u$. *Stretching the utterance world* in order to cover different temporal positions makes new backgrounds of information relevant to evaluating the content of that utterance and allows for a principled explanation of the transition from a definite truth value to its opposite, even if this kindof transition concerns a *restricted* class of utterances, namely the utterances expressing a future-oriented proposition.

As for Evans' criticism, the kind of variability discussed here is at the same time restrained (because it affects only the contents of a circumscribed type of utterances, i.e., the utterances involving a reference to future courses of events) and systematic (for it is not confined to the transition from indefinite to definite truth-values, but allows for transitions from truth to falsehood and vice versa).

12. Time, Tense and Contexts

This is just a provisional conclusion, for a more careful account of the role of time in fixing the relevant truth-conditions is in order at this point.

As an illustration, consider two possible utterances of a by now familiar example, repeated here as (26):

(26) Bill Evans is playing with Jim Hall.

In the first scenario, (26) is a very natural answer to a question (concerning the identity of Bill Evans' partner) asked by a person during a concert at the

Montreux Jazz Festival. What is involved here is an event *which is occurring at the utterance moment*.

But, as we saw above, (26) can be used in a different context, in order to speak of a *planned* event,whose occurrence is located in the *future* of the present moment.

Now the question is: how is time involved in determining the appropriate truth-conditions in such scenarios? At least two roles can be detected here. One of them is quite familiar: when we speak of the *evaluation time*, we mean for instance that by uttering (26) at time *u*, in the first scenario, the speaker says something true because *at that very moment* there is an event in progress of Bill Evans' playing with Jim Hall. In this case the evaluation time coincides with the utterance time.

But there is also *the time we are speaking about*, which coincides with the utterance time (and the evaluation time) in the first scenario, but not in the second scenario, where the situation is more complex: once more, evaluation, time, and utterance time coincide (for it is at this very moment that we want to judge the statement at issue as true or false, if it is used, for instance, as an answer to a question like "Is it true that ...?"), but they do *not* coincide with the time which is spoken about (the time of the tomorrow concert).

This kind of implicit reference can be fixed by contextual factors such as a previous discourse (in the case of an anaphoric link, as suggested by the second scenario) or current evidence (our presence at the concert, in the first scenario). Intuitively speaking, the idea is that an utterance of a sentence like (26) *concerns* a particular temporal situation, which can be located in the present, the past or the future of the utterance moment. This *time which is spoken about*[28] (a point or interval in a branching structure, according to the formal framework adopted here) has a crucial role to play indefining the content of an utterance.

This is the role Frege has in mind when in a famous passage he explains how the utterance time contributes to determining the time we refer to by using a tensed sentence: "If a time indication is needed by the present tense, one must know when the sentence was uttered to apprehend the thought correctly. Therefore, *the time of utterance is part of the expression of the thought*. If someone wants to say the same today as he expressed yesterday using the word "today", he must replace this word by "yesterday" [...] The mere wording, as it is given in writing, is not the complete expression of the thought, but the knowledge of certain accompanying conditions of utterance, *which are used as a means of expressing the thought*, are needed for its correct apprehension" (Frege 1918: 24; italics mine.) Thus, the "complete" expression of a thought or proposition must contain a specification of the time the statement at issue is about, and thanks to such a specification (made possible by the—possibly implicit—reference to the

[28] This is *the time we aim at* in order to locate an event *from a given perspective point*, which is also temporally located.For the present purposes there is no need to make this notion more precise, e.g. by resorting to the classical distinction between event time and reference time (Reichenbach). Klein defines the *topic time* as "the time span to which the speaker's claim is confined" (Klein 1994: 4). In Bonomi 2010, I talk of a *target time*, by resorting to a metaphorical distinction between an aiming device and the target aimed at by that device. In what follows, I use the generic term "time which is spoken about" to avoid a theoretical commitment which is not required in the present context.

utterance moment) the evaluation of the thought or proposition at issue is fixed once and for all. And the stability of evaluation, to use Evans' wording, follows.

The contrast, here, is between a complete expression of the thought or proposition and an incomplete one. However, and this is the characterizing feature of eternalism, the latter has *no* semantic relevance. There is no intermediate entity, namely a temporally *neutral* proposition, which accounts for the dependency of evaluation on a temporal parameter. This is so for the simple reason that such a parameter is *incorporated* into the expression of the thought.

I will not address here Kaplan's well-known argument against this line of thought, an argument based on the role of temporal operators: applying these operators, so runs the objection, to propositions where the temporal information is completely specified would be tantamount to using them vacuously.[29] I will turn instead to the role that Prior attributes to temporally neutral propositions to account for some peculiar uses of tensed sentences.

Interestingly enough, his starting point is the same as Frege's: the time a proposition *is about* (which, in many cases, coincides with the utterance time) is an essential ingredient to determine the full content expressed by an utterance event:

> [A tensed language] *implicitly refers* to the time of utterance, and by tensing what is implicitly said of the time of utterance it can indirectly characterise other times also [...] In at least the most elementary tensed languages instants or times are not mentioned, but tensed propositions are understood as directly or indirectly characterising the *un*mentioned time of utterance (Prior and Fine 1977: 30).

So, on this account, the time which is spoken about, with its anchoring effect, plays a crucial role in determining the full content expressed by an utterance. Still, we can insist that there are plausible reasons to isolate a notion of content which is *independent* of that kind of anchoring.

To see this, consider the following situation.[30] On November 27, 2011, Leo, a famous economist, says in an interview:

(27) Italy is facing a severe crisis.

As everyone knows, this a true statement. One year later, after reading the old interview, he comments:

(28) Thank Goodness, what I said one year ago is no longer true. (Italy is out of the crisis.)

Now, consider (27) and suppose that, as required by the kind of temporal anchoring suggested by Evans in order to get "eternal" propositions, the time which is spoken about (*and which coincides with the utterance time in this case*) is

[29] See Recanati 2007 for a reconstruction of the debate between eternalists and temporalists.

[30] This example is reminiscent of Prior's "Thank Goodness, it's over". Notice that the situation depicted by a sequence like (27)-(28) can be more complex. Imagine two economists, *A* and *B*, who speak different languages. For example, one of them utters (27), whilst the other, who speaks Italian, utters "L'Italia sta attraversando una crisi molto seria". Supposing that these utterance events take place at the same time, one year later an observer *C* might comment: "Thank Goodness, what *A* and *B* said one year ago is no longer true".

incorporated into the content expressed by Leo's utterance. If the expression "what I said one year ago" refers to this kind of content, by uttering (29) Leo states something absurd, because, under this assumption, what he means is that it is no longer true that Italy was facing a severe crisis on November 27, 2011.

If we look at the content expressed by Leo when he utters (27) to speak of the Italian crisis, weobserve the following:

(Profile 1) utterance time = evaluation time = time which is spoken about (*TT*).

It is the utterance time that Leo has in mind when he utters (27) in order to locate the relevant eventor state (Italy's crisis) and it is with respect to this very moment that his utterance is to be evaluatedas true or false. But if it is true, of the utterance moment *u*, that Italy is facing a severe crisis, then there is *no* moment *t*, such that: *t* □ *u* and it is false at *t* that Italy is facing a severe crisis at *u*. No variability of truth value over time is admissible if we stick to the original utterance time as the temporal situation Leo's statement refers to: which means that there is no way to explain why (28) does make sense.

Indeed, the comment made by uttering (28) can be plausible (and true) only by associating the expression "what I said" to a proposition which is *not* anchored to the utterance time of (27), and which includes a *shiftable* component. In other terms, we have to isolate a temporally neutral content that can be obtained by abstracting over the parameter represented by the utterance time of 27, which coincides with the evaluation time and with *TT*.

The point is that, by uttering (28), Leo does *not* intend to *revise* his original statement, which was, is and will be true: the expression "what I said', in (28), denotes a content that is not temporally anchored to the time which is spoken about in (27), i.e., the utterance time of (27). More exactly, such a content can only be obtained by *abstracting over* that contextual parameter (which coincides with *TT*), for there is a sense in which (28) might be paraphrased as follows:

(28') If I should now say what I said one year ago, I would say something false

where the propositional content referred to by the expression "what I said" is not anchored to the situation Leo had in mind when he uttered (27).

Similar remarks apply to the Sarah Palin's case discussed above. But there are some interesting differences. When, on June 28, Leo utters:

(29) A woman will run for President

he says something intuitively true. But, on October 28, after Sarah Palin's withdrawal and Michael Moore's nomination, he mightcomment:

(30) What I said three months ago is no longer true. (A man will run for President.)

There is a strong similarity, of course, between (28) and (30), because both raise a problem of truth-value variability (whatever you think of this issue). But the Sarah Palin story has a peculiarity which deserves a short reflection. What distinguishes the sequence (29)-(30) from (27)-(28) is that the expression "what I said" in (30) denotes a content temporally *anchored* to the time which is spoken about in (29): as shown by the second sentence in (30), the speaker is still referring to the time of the next Presidential election.

What he means, by uttering (30), is that the *anchored* proposition associated to the utterance of (29) is no longer true at the new evaluation time. In other

terms, the evaluation time, which coincides with the utterance time, is made shiftable by abstraction, but the time which is spoken about (i.e., the time of the Presidential Election) remains *unchanged*.

The difference, with respect to (27)-(28), is that (29), unlike (27), has the following profile:

(Profile 2) utterance time = evaluation time □ time which is spoken about (*TT*).

That is why we can speak of a *revisable* statement made by uttering (29): what the speaker said *about* a given time located in the future is judged to be true at the evaluation time *t*, but no longer true at the evaluation time t′.

This peculiarity of future-oriented statements comes as no surprise in the theoretical framework adopted here: the passing of time modifies not only the state of the world, but also the state of the relevant information, which is an essential ingredient of the truth-conditions for this kind of statements.

To conclude these informal remarks about the variability of truth values, let us sum up the mainpoints of the above discussion.

First of all, there is *the time which is spoken about* (*TT*) when a tensed sentence is uttered. Forexample, we have just seen that in the case of (26), repeated here

(26) Bill Evans is playing with Jim Hall,

depending on the context, the time at issue can be the utterance time (if, for example, the speaker intends to identify Evans' partner on the stage during a concert) or a future time (if she is speaking of the tomorrow concert). In general, the content which can be associated to an utterance event can be seen as a content *anchored* to the relevant *TT* or *independent* of it. And we have considered, withPrior, the need for the second kind of content (temporally neutral propositions) in order to account for the feeling of relief expressed thanks to a statement like (28) by a speaker *located in time*. To go back to a familiar example, saying that the proposition associated to an utterance of a sentence like "Socrates is sitting" can be true at time *t*, but false at time *t′* is just a way to recall us that there are situations in which it can be relevant to isolate what remains of an anchored content once a contextual feature has been stripped off. If *X* says "Socrates is sitting" at moment *t*, and *Y* says "Socrates is sitting" at moment *t′*, there is a sense in which we can state that they say the same thing, but there is also a sense in which we can state that they say different things, for the simple reason that *the times which are spoken about are different*.

Stability of evaluation (in terms of truth-values) may be guaranteed by keeping *TT* fixed. Indeed, this kind of anchoring puts severe restrictions on the role of the *evaluation time*. To see this, suppose, for instance, that someone asks Leo what Lea did yesterday at 3 p.m. and that he answers:

(A) She went to the doctor.

Now, if yesterday at 3 p.m. Lea went to the doctor's and if what Leo says is anchored to therelevant *TT* (yesterday, 3 p.m.), it is quite natural to suggest that what he says is true at the utterance time *u* and at *any* evaluation time *t,* such that *u* □ *t*. And the same holds of "Socrates is sitting', once the content has been *anchored* to the time which is spoken about (and which coincideswith the utterance time). Changing the utterance time (and, as a consequence, changing the evaluation time) has no effect *if the time which is spoken about remains unchanged*.

This is true of statements in the past or present tense.[31] But what about future-oriented statements? What may happen, in such cases, is that *although TT* is kept unchanged, different truth values can be associated to different evaluation times, witness a statement like:

(28) It is no longer true that Bill Evans is playing with Jim Hall [in the tomorrow concert].

To sum up, there is a first level of truth-value variability: it concerns the content of an utterance when this content is individuated *independently* of the time which is spoken about (and which in many cases coincides with the utterance time). We might speak of "floating" propositions in such cases, and they can have a theoretical role to play, for instance, in order to account for intentional states of mind.[32] At a second level of analysis there are "anchored" propositions. They can be seen as ordered pairs consisting of a proposition of the first type *and* the time which is spoken about: *their evaluation is stable*, *unless their anchoring involves a future time*. This means that, unlike other types of statements, future-oriented statements, at least in some cases, are *revisable*, for theyinvolve truth-value variability in a deeper sense: due to the non-monotonicity of the sequence of relevant states of information, the anchored proposition we get in such cases by keeping *TT* constant may turn out to be true (false) at time t, but false (true) at time t', where $t < t'$.

This kind of revisability raises an interesting problem of theoretical adequacy, for a complete context of utterance (where all the necessary indexical information is specified) makes a plurality of evaluation times relevant to define the notion of truth in *that* utterance context. The point is that,unlike Kaplan's framework, the kind of analysis developed here does *not* associate a context of utterance to a *single* evaluation moment in order to define the notion of truth in context, since the utterance event is seen as belonging to a world in progress where possibly different states of information may follow each other as times goes by. And since these states of information can be associated to different moments in a non-monotonic way, truth-value variability follows.

As specified above, given a context $c = <u, TT_c, VIEW_u>$, the dynamic side of this context is represented by the *function* $VIEW_u$, which maps moments to states of information. More exactly, forany moment t, $VIEW_u(t)$ is the background of information which is relevant at t.

So, on this proposal, contextual dependency manifests itself in a twofold way:

[31] "One of the big differences between the past and the future is that once something has become past, it is, as it were, out of our reach—once a thing has happened, nothing we can do can make it not to have happened. But the future is to some extent, even though it is only to a very small extent, something we can make for ourselves. And this is a distinction which a tenseless logic is unable to express. In my own logic with tenses I would express it this way: We can lay it down as a law that whatever *now is* the case *will always have been* the case; but we can't interchange past and future here and lay it down that whatever *now is* the case *has always been going to be* the case—I don't think that's a logical law at all" (Prior 1996).

[32] Indeed, a philosophical justification for this two-layered analysis might be the following: propositions are intentional entities involving an object (the time which is spoken about), and they can be considered independently of or in relation with such an object.

(i) by narrowing down the location of the time span which *is spoken about* (TT_c);
(ii) by narrowing down (thanks to *VIEW*) the temporal location at which the relevant background of information must be associated.

Given a context *c*, by abstracting from the time fixed in (i), which often coincides with the utterancetime, one gets "floating" propositions, whose theoretical relevance has been proposed by temporalists à la Prior. But even when an "anchored" proposition is determined by keeping *TT* fixed, truth-value variability is possible, because of the functional nature of $VIEW_u$. From an intuitive point of view, this means that a *plurality* of temporal situations, instead of a single one, is available to define the notion of truth in *that* context.

This is why a person who, at time *u*, utters a future-oriented sentence like (26) is prepared to revise her statement at a time $t > u$, in the presence of a new background of information. To put it in a slightly different way, it might also be said that there is here a sort of *announced indeterminacy* as concerns the evaluation time which is selected by the utterance context or that a plurality of contexts should be associated to the utterance event at issue.

13. Conclusions

a. Towards a Cartography of Propositional Contents

In the theoretical framework under discussion different ways of determining the content that can beassociated to an utterance are available. To see this in a simplified form, let us assume that function $VIEW_u$ (which fixes a relevant background of information for any evaluation moment) is *implicitly* provided by the context (see Sect. 10 for the explicit version), so that a context *c* is a pair of moments $<u,v>$, where moment *u* is the utterance time and moment *v* is the time which is spoken about.

It should be kept in mind that, as argued in Sect. 9, on this approach moments in the tree can be seen as situations or world-states (Prior), which represent different alternatives both in a temporal grid *and* in a logical space. Anyway, in order to preserve a more familiar terminology, I will continue to use expressions like "utterance *time*" or "evaluation *time*".

To go back to our examples, with this simplification in mind the evaluation of (30) ("A woman willrun for president"), relative to the two scenarios described above, can be stated as follows:

$[\![(30)]\!]^{<u,x>,g,u} = 1$ where $u = sit_1$ and $g(x) = sit_3$ (presupposition)
$[\![(30)]\!]^{<u',x>,g,u'} = 0$ where $u' = sit_2$ and $g(x) = sit_3$ (presupposition)

Here sit_1, which plays the role of utterance time and evaluation time in the first scenario, is the temporal situation corresponding to the first Republican Convention. sit_2, which plays the role of utterance time and evaluation time in the second scenario, is the temporal situation corresponding to the second Republican Convention. sit_3 (the time which is spoken about in both scenarios) is the temporal situation corresponding to the intended Presidential Election.

As we saw, the idea is that *TT* does not change when passing from the first utterance context to the second one (in both cases it is the time of the next Presidential Election). This is why the expression "what I said", in (30), denotes a content *temporally anchored to the time which is spoken about* in (29): as shown by

the second sentence in (30), the speaker is *still* referring to the time of the next Presidential Election. What he means, by uttering (30), is that the *anchored* proposition associated to the utterance of (29) is no longer true at the new evaluation time. In other terms, the evaluation time (which coincides with the utterance time) is shifted, but *TT* remains *unchanged.*

This means that if, by *lambda abstraction*, we want to determine an appropriate content, the value which should be assigned to variable *x* is NOT shiftable and that, unlike the utterance time (which coincides with the evaluation time), it cannot be □-bound. So, what we get is the following proposition:

□□□□□ $\nu[\![(29)]\!]^{<v,x>,g,v}$ $g(x) = sit_3$ (presupposition)

which, applied to the relevant situations, yields the intended result:

□ $\nu[\![(29)]\!]^{<v,x>,g,v}$ $(sit_1) = 1$ $g(x) = sit_3$ (presupposition)
□ $\nu[\![(29)]\!]^{<v,x>,g,v}$ $(sit_2) = 0$ $g(x) = sit_3$ (presupposition)

We also noticed that the case of the economist's example is different: the time which is spoken about *does change* when passing from the first utterance context to the second one, for it coincides with the utterance time. So, the appropriate content can be represented as the following proposition,where the time which is spoken about is shiftable (is □-bound):

□□□□□ $\nu[\![(27)]\!]^{<v,v>,g,v}$

Indeed, as shown by the discussion about (27)-(28), the comment made by uttering (28) ["What I saidone year ago is no longer true"] can be plausible (and true) only by associating the expression "what I said" to a proposition which is *not* anchored to the time which is originally spoken about and which includes a *shiftable* component. We have here a *temporally neutral* content that can be applied to different temporal situations. This proposition, applied to the relevant temporal situations, yields the intended result:

□ $\nu[\![(27)]\!]^{<v,v>,g,v}$ $(sit_1) = 1$
□ $\nu[\![(27)]\!]^{<v,v>,g,v}$ $(sit_2) = 0$

where sit_1 is the temporal situation corresponding to the interview, sit_2 is the temporal situation in which the economist comments his old statement.

The proposition in (B) is the kind of content (associated to the utterance of (27)) that the temporalist proposes in order to account for the comment made by uttering (28).

But what happens if we *stick to the time which is originally spoken about* when (27) is uttered (and which coincides with the utterance time and the evaluation time)? In this case no parameter is abstracted over and what we get is *eternalism*:

(C) □ $\nu[\![(27)]\!]^{<x,x>,g,x}$

where $g(x)$ = the situation corresponding to the utterance time (presupposition).

This is a constant function (proposition), since, for any situation *s*, such that $g(x)$ □ *s*:

□ $\nu[\![(27)]\!]^{<x,x>,g,x}$ $(s) = 1$

where g(x) = sit_1 (the temporal situation corresponding to the interview).

By sticking to the time which is originally spoken about (= the utterance time) we get a proposition whose truth value is fixed once and for all, for any *s* it applies to. The moral we can draw is that different propositional contents are available here, depending on the different scenarios.

If, for instance, we are interested in the correctness of the Leo's utterance *with respect to the situation he referred to*, (C) is the content we should appeal to: as we have just said, it is a constantproposition which yields the same truth value at any time (starting from the utterance time). On the contrary, (B) is the natural candidate if, by *abstracting from* the time which is originally spoken about (and which coincides with the utterance time), we look at the intentional states of agents *located in time*. This is what justifies the feeling of *relief* associated to the utterance of (29), exactly as in Prior's original Thank-Goodness example. As for the kind of proposition sketched in (A), it occupies an intermediate position, for *TT* stays fixed, while the utterance time (which coincides with the evaluation time, but not with the time which is spoken about) is abstracted over. This is the case of future-oriented statements like (26) and (29), whose evaluation depends on *changing* backgrounds of information.

To sum up:

(i) The examples involving Sarah Palin and Bill Evans show that, *even by keeping the TT parameter fixed* (condition *S*) there are contents of utterances (propositions) of type (A) which can change truth value over time. Itis the case of (some) future-oriented statements, which illustrate a *first type of non-persistence*.

(ii) In the case of past (or present) tensed sentences it is possible to get "non-persistent" propositions of type (B) *only* by relaxing condition *S*. This means that we have to abstract over the *TT* parameter by identifying it with a varying utterance time: what we get is a *second type of non-persistence*.

b. A First Look

It is not in the scope of the present paper to start a systematic scrutiny of the propositional profiles that can be individuated along these lines. A preliminary look might be instructive. Let us start with familiar cases.

1. "Eternal" Propositions (Constant Functions): $\square\, \nu[\![\square]\!]^{<x,x>,g,x}$ (= case C above)

The truth value is fixed once and for all, independently of the temporal situation this function applies to (starting from the utterance moment), that is

$\square\, \nu[\![\square]\!]^{<x,x>,g,x}(t) = \square\, \nu[\![\square]\!]^{<x,x>,g,x}(t')$ for any t and $t' \square\ g(x)$.

Example (discussed above): $\square$ is the sentence "Italy is facing a severe crisis", and $g(x)$ = the time at which Leo utters this sentence.

Possible comment (one year later): I've just checked all the relevant data. *What Leo said* last year is (was) true: Italy was really facing a severe crisis at that time. (Notice that both the present tense and the past tense can be associated to the truth predicate.)

2. (Totally) "Diagonal" Propositions: $\square\, \nu[\![\square]\!]^{<v,v>,g,v}$ (= case B above)

The utterance moment coincides with the time which is spoken about and with the evaluation time.We can have different truth values at different times.

Example (discussed above): $\square$ is the sentence "Italy is facing a severe crisis".

Possible comments: Thank Goodness, *what Leo said* one year ago is *no longer* true.(Italy is out of the crisis.) Or: If Leo should now say what he said one year ago, he would say something false.

Question: since in such cases phase adverbs like "no longer" or "still" may involve propositions that are obtained by abstracting over contextual parameters, should their behaviour be qualified as "monstrous" according to Kaplanian standards?

3. (Partially) "Diagonal" Propositions: □ $\nu[\![\square]\!]^{<\nu,z>,g,\nu}$ (= case A above)

The utterance moment coincides with the evaluation time but not with the time which is spoken about (*TT*). *TT* is contextually fixed. We can have different truth-values at different moments (even if the time which is spoken about does *not* change.) In particular, this is the case of future-oriented statements.

Example (discussed above): □ is the sentence "A woman will run for President".

Possible comment (on October 28): *What Leo said* after the first Convention is *no longer* true. (The person who will run for President is *no longer* a woman.)

Question: the same raised in the case of totally diagonal propositions.

4. Variable Evaluation Time: □ $\nu[\![\square]\!]^{<u,z>,g,\nu}$

The utterance time and TT are contextually fixed. We can have different truth-values at different moments (even if the time which is spoken about does *not* change).

Example: □ is the sentence "There will be a sea-battle". (As an answer to the question: What will happen tomorrow at 3 p.m.?) *TT* = tomorrow, 3 p. m. According to a familiar interpretation (MacFarlane 2003, 2008), this proposition is neither true nor false when evaluated at the utterance time itself, but can be evaluated as true at a later moment (e.g., when the battle has just started).

A possible comment (during the battle): *What Leo said* yesterday was true.

5. Variable TT: □ $\nu[\![\square]\!]^{<u,\nu>,g,\nu}$

The time which is spoken about (and which coincides with the evaluation time) is abstracted over.We can have different truth values at different moments.

Example: imagine that Leo, the computer engineer of our department, has just found out that the anti-virus system in the LAN of the Students Room is not regularly updated. There are situations in which the system is updated, but in other situations it isn't. So, he gathers all the students and says:

(A) Connecting a computer might be dangerous here.

Now suppose that at moment *t* the system is updated and that Lea knows that. Thus, she connects her computer because she knows that (A) *is not true in that particular situation*, i.e., at moment *t*. But suppose also that at moment *t'* ($t < t'$) the system is *not* updated and that Theo knows that. As a consequence, he does not connect his computer for he knows that (A) *is true in that particular situation*.

Lea's comment: I'm lucky. *What Leo said* is not true in my case.

6. Variable Evaluation Time: □ $\nu[\![\square]\!]^{<z,z>,g,\nu}$

TT coincides with the utterance time.

Example: □ is the sentence "Italy is facing a severe crisis", where *TT* (which coincides with the utterance time) is contextually fixed. The evaluation does not change for any ν such that $g(z)$ □ ν.

A possible comment: I've just checked out the data. *What Leo said* one year ago [about that situation] is (was) true.

7. Another Kind of "Eternal": □ $\nu[\![\square]\!]^{<z,x>,g,z}$

The evaluation time coincides with the utterance time but not with *TT*.

Example (of the type discussed in Partee 1973, assuming that *TT* is in the past of the utterance time): □ is the sentence "I turned on the alarm system" (as an answer to the question "What did you do when you left?"). The truth-value is fixed once and for all at the utterance moment z and does not change across time.

A possible comment: I've just checked the videotape. *What Leo said* is true.

8. Variable TT: □ $\nu[\![\square]\!]^{<z,\nu>,g,z}$

The time which is spoken about is abstracted over whilst the other contextual parameters stay fixed.

As an example, consider the following exchange (original *TT* = this week).

A: What did you do this week for the course of logic?

B: I proved at least five theorems in the exercises' booklet.

A: I don't think so. *What you said* is true of the last week, not of this week.

The moral to be drawn, after this short excursus, is that there is no propositional profile (be it ascribable to eternalism or temporalism, or whatever) that can be associated to utterances *in general*, because:

(i) there are kinds of utterances that can be associated to some kinds of propositional profile but not to others (think, for instance, of the contrast between past-oriented and future-oriented statements);

(ii) the same utterance can be associated to different propositional profiles, depending on whichcontextual parameters stay fixed and which are abstracted over.

This is clear, as we have just seen, in the case of the utterance of (27) discussed above:

(27) Italy is facing a severe crisis.

Indeed, this utterance event can be associated to an "eternal" proposition (of type □ $\nu[\![\square]\!]^{<x,x>,g,x}$) if we are concerned with what Leo, the famous economist, said *about a given temporal situation* (November 2011). But it can also be associated to a non-persistent proposition (of type □ $\nu[\![\square]\!]^{<\nu,\nu>,g,\nu}$) if we abstract from (by abstracting over...) that temporal location and we focus on alternative time spans, as shown by the comment "Thank Goodness...".

In these cases, multiple propositional contents are available for the same utterance event, since *what we abstract from* in order to determine *what was said* depends on the conversational situations we are engaged in when we talk about that event.

References

Abusch, D. 2012, "Circumstantial and Temporal Dependence in Counterfactual Modals", *Natural Language Semantics*.

Bonomi, A. 1997, "The Progressive and the Structure of Events', *Journal of Semantics*, 14, 173-205.

Bonomi, A. 2006, "Truth in Context", *Journal of Semantics*, 23, 107-34.

Bonomi, A. 2010, *Imperfect Propositions*, http://filosofia.dipafilo.unimi.it/bonomi/impbonomi.pdf.

Bonomi, A. & Del Prete, F. 2008, "Evaluating Future-Tensed Sentences in Changing Contexts", Unpublished manuscript, Università degli Studi di Milano, http://www.filosofia.unimi.it/~bonomi/future.pdf

Belnap, N., Perloff, M., and Xu, M. 2001, *Facing the Future: Agents and Choices in Our Indeterminist World*, Oxford: Oxford University Press.

Condoravdi, C. 2001, "Temporal Interpretation of Modals", in Beaver, D., Kaufmann, S., Clark, B., and Casillas, L. (eds.), *Stanford Papers on Semantics*, Palo Alto: CSLI Publications.

Copley, B. 2009, *The Semantics of the Future*, New York: Routledge.

Del Prete, F. 2010, *Non-Monotonic Futures*, PALMYRIX: Logic and the Use of Language, ILLC, Amsterdam; Institut Jean-Nicod, Paris, Jun 2010, Amsterdam, Netherlands, hal-01408355

Dowty, D. 1979, *Word Meaning and Montague Grammar*, Dordrecht: Reidel.

Dummett, M. 2004, *Truth and the Past*, New York: Columbia University Press.

Evans, G. 1985 "Does Tense Logic Rest on a Mistake?", in *Collected Papers*, Oxford: Clarendon Press, 341-63.

Frege, G. 1918, "The Thought: A Logical Inquiry", Engl. Transl. 1956 by A.M. Quinton and M. Quinton, *Mind*, 65, 289-311.

Grove, A. 1988, "Two Modelling for Theory Change', *Journal of Philosophical Logic*, 17, 155-70.

Heim, I. 1994, "Comments on Abusch's Theory of Tense", in Kamp, H. (ed.), *Ellipsis, Tense and Questions*, University of Amsterdam, DYANA Deliverable R.2.2.B, 143-70.

Kaplan, D. 1977, "Demonstratives", repr. in Almog, J., Perry, J., and Wettstein, H., *Themes from Kaplan*, Oxford: Oxford University Press, 1989.

Katz, G. 2003, "Event Arguments, Adverb Selection, and the Stative Adverb Gap", in Lang, E., Maienborn, C., and Fabricius-Hansen, C. (eds.), *Modifying Adjuncts*, Berlin: Mouton de Gruyter.

Klein, W. 1994, *Time in Language*, London and New York: Routledge.

Kölbel, M. 2009, "On Future Contingents", paper delivered at the Conference on *Language and Temporality*, Università dell'Aquila, September 8-10, 2009.

Kratzer, A. 2012, *Modals and Conditionals*, Oxford: Oxford University Press.

Krifka, M. 2000, "Alternatives for Aspectual Particles", presented at the 26th Meeting of the Berkeley Linguistics Society.

Lewis, D. 1973, *Counterfactuals*, Cambridge, MA: Harvard University Press,

Lewis, D., 1986, *On the Plurality of Worlds*, Oxford: Basil Blackwell.

MacFarlane, J. 2003, "Future Contingents and Relative Truth", *The Philosophical Quarterly*, 53, 321-36.

MacFarlane, J. 2008, "Truth in the Garden of Forking Paths", in Kölbel, M. and García-Carpintero, M. (eds.), *Relative Truth*, Oxford: Oxford University Press.

Mondadori F., 1978, "Remarks on Tense and Mood: The Perfect Future", in Guenthner, F. and Rohrer, C. (eds.), *Studies in Formal Semantics*, Amsterdam: North-Holland, 223-48.

Partee, B. 1973, "Some Structural Analogies Between Tenses and Pronouns in English", *Journal of Philosophy*, 70, 601-609.

Øhrstrøm, P. and Hasle, F.V. 1995, *Temporal Logic*, Dordrecht: Kluwer.

Prior, A.N. 1957, *Time and Modality*, Oxford: Clarendon Press.

Prior, A.N. 1967, *Past, Present and Future*, Oxford: Clarendon Press.

Prior, A.N. 1968, *Papers on Time and Tense*, Oxford: Clarendon Press.

Prior, A.N. 1996, "Some Free Thinking about Time", in Copeland, B.J. (ed.), *Logic and Reality: Essays on the Legacy of Arthur Prior*, Oxford: Clarendon Press.

Prior, A.N. and Fine, K. 1977, *Worlds, Times and Selves*, London: Duckworth.

Recanati, F. 2007, *Perspectival Thought*, Oxford: Oxford University Press.

Thomason, R.H. 1970, "Indeterminist Time and Truth Value Gaps", *Theoria*, 36, 264-81.

Thomason, R.H. 1984, "Combinations of Tense and Modality", in Gabbay, D. and Guenthner, F. (eds.), *Extensions of Classical Logic*, *Handbook of Philosophical Logic*, Vol. II, Dordrecht: Reidel: 205-34.

Argumenta 9, 1 (2023)
Critical Discussion

On Patrick Todd's *The Open Future: Why Future Contingents Are All False* OUP 2021

The Journal of the Italian Society for Analytic Philosophy

Future Contingents, Open Futurism, and Ontic Indeterminacy

Giuseppe Spolaore

University of Padua

Abstract

This paper critically discusses Patrick Todd's book, *The Open Future: Why Future Contingents Are All False* (Oxford: Oxford University Press, 2021).

Keywords: Future contingents, Peirceanism, Presentism, Semantics, Future tense.

It is February 2022. Russia has not yet invaded Ukraine, but (let us assume) the invasion is not inevitable. You and I disagree on whether it will take place. You claim:

(1) Russia will invade Ukraine in days.

In support of (1), you cite respectable intelligence sources. I strongly reject your claim:

(2) Russia will not invade Ukraine anytime soon.

A few days later, Russia begins its full invasion of Ukraine. At this point, we can say in retrospect that your prediction, (1), was true and mine, (2), false. Can we?

According to Patrick Todd's *The Open Future*, we cannot. In fact, both predictions were false from the very start. And of course, the problem is not limited to these claims of ours. Any *future contingent* (any prediction of future events that are neither inevitable nor impossible at speech time) is bound to be false:

Open future view (OF). All future contingents are false.

Of course, (OF) sounds crazy. But in his book, Todd makes an excellent case for the conclusion that, if we take two reasonably popular philosophical doctrines seriously enough, then (OF) becomes attractive. These are *causal indeterminism* (roughly, the view that the totality of present events and the laws of nature do not determine a unique future) and *no-futurism* (the A-theoretical view that no future thing exists). More specifically, Todd assumes *presentism* (the view that only present things exist).

The book is divided into an introduction and eight chapters. Here, I will mainly cover some key aspects of the introduction and chapters 1-3, in which the bulk of Todd's view is presented, and I will briefly discuss the content of chapters

Argumenta 9, 2 (2023): 159-172
ISSN 2465-2334

© 2023 Giuseppe Spolaore
DOI 10.14275/2465-2334/202317.spo

4, 6, and 8, which deal with specific problems of the view. I will ignore chapters 5 (*Omniscience and the Future*) and 7 (*Future Contingents and the Logic of Temporal Omniscience*), as they concern theological issues on which I have little to say.

At the heart of the book lies the so-called *problem of future contingents*, which Todd briefly discusses in the introduction. A *prima facie* strong argument, reported by Aristotle in the *De Interpretatione*, seems to directly lead us from the principle of bivalence (the view that every statement is either true or false) to fatalism (the strongly counterintuitive view that everything is forced upon us as a matter of logico-semantical principles). Consider the following claim, assuming it is uttered today:

(3) There will be a sea-battle tomorrow.

By bivalence, (3) is either true or false now. Suppose it is true now. If so, then its truth is inevitable (it is always too late to change what holds *now*). Similarly, if (3) is false now, then its falsity is inevitable. Thus, (3) is now either necessarily true or necessarily false. But there is nothing special in (3). Therefore, everything future is forced upon us: fatalism is true.

Before introducing Todd's solution to the problem, let me lay out my cards. My favorite solution posits an ambiguity between two readings of time-relative truth ascriptions such as:

(4) Statement (3) is true now.

In the first reading, what we mean is that the statement is true if interpreted/evaluated relative to that time (the expression "interpreted/evalued" actually hides another ambiguity, which for our current purposes can be ignored). In the second reading, what we mean is that the statement is *historically necessary* at the relevant time. This reading is easily recognizable because it allows us to add aspectual clarifications that are often misplaced in other readings. For instance, we can restate (4), in this second reading, by saying that (3) is *already* true (now). In this sense, a claim like "It is *now* true that the Giants will win the finals" conveys the same thought we can express by saying [before the finals] "The Giants have *already won* the finals". Either way, we are saying or at least strongly implicating that the Giants' win is already inevitable at speech time. This is the reason why there is no Moorean infelicity in the following claim:

(5) It is not *now* true that the Giants will win the finals, but I believe they will win the finals.

If I am right, the problem with Aristotle's argument for fatalism is that it subtly equivocates between these two readings of (4).[1]

When a B-theoretical, tenseless conception of truth and reality is adopted, this solution allows to keep the principle of bivalence without endorsing fatalism. These are considerable advantages. Todd's favorite solution shares the same advantages, but assumes an A-theoretical, tensed conception of truth and reality. The price is a rather radical departure from a variety of common intuitions concerning future-tensed statements and their truth conditions. As already mentioned, Todd subscribes to (OF): all future contingents, including (3), are false.

[1] See Tooley 1997 and Spolaore and Del Prete 2019. This solution to the problem of future contingents can be laid down either in a B-theoretical frame, by appeal to a fundamentally tenseless notion of truth (this is my favourite approach) or in a more A-theoretically friendly way, e.g., by appeal to MacFarlane's (2003) distinction between context of utterance and context of assessment.

Thus, their negations are true, and both the principles of bivalence and excluded middle are preserved. Fatalism is avoided by adopting a non-standard, but pretty natural conception of indeterminacy, which I will introduce in due course.

Chapter 1 (*Grounding the Open Future*) concerns the relations between presentism and (OF). Presentists adopt a symmetrical stance towards the past and the future: neither of them exists in any sense (there are no past/future things, facts, events). As a result, some key arguments for the openness of the future, if stated in a presentist framework, also support the symmetrical view that the *past* is open—let us call it (OP). To exemplify, Todd considers an argument for (OF) based on the principle that *truth supervenes on reality* (TSR):

> 1. If there are true future contingents, the truth of these future contingents would not supervene on present reality [present events and the laws of nature].
> 2. But all truth supervenes on present reality.
> So, 3. There are no true future contingents (12).

This argument for (OF) sounds plausible, but it equally supports (OP). For if present events and the laws of nature are compatible with more than one possible future, then they are equally compatible with more than one possible past (physical laws are time reversible). Thus, presentists cannot adopt (OF) based on the argument from TSR, unless they are also prepared to accept (OP). In Todd's mind, however, (OP) is unpalatable. A key aim of this chapter is to reject the argument from TSR and, in its place, provide an argument for (OF) that does not support (OP). Against TSR, Todd observes that the following counterfactual claim is plausible.

> (SBP) If it is true that there was a sea-battle in 2019, it would still be true that there was a sea-battle in 2019, even if everything went out of existence, and there came to be nothing at all (15).

As a clarification about (SBP), he adds:

> What I mean is that if everything on which the truth of this proposition could plausibly be thought to supervene went out of existence, the claim
> [(6) There was a sea-battle in 2019]
> would still be true (16).

But if truths about the past do not depend on their (present) supervenience base, Todd argues, then TSR is not generally valid. Some truths, most noticeably, truths about the past, are *brute*: they remain unaffected if we remove their supervenience base. In contrast, Todd notes, the future-directed analogue of (SBP) sounds implausible:

> (SBF) If it is true that there will be a sea-battle in 2219, it would still be true that there will be a sea-battle in 2219, even if everything went out of existence, and there came to be nothing at all (20).

Todd concludes that truths about the future *do* depend on their supervenience base. Thus, the argument from TSR is only correct when applied to truths about

the future. Truths about the past are brute, they do not depend on their supervenience base. Therefore, the argument from TSR brings no support to (OP).

Todd's reasoning is ingenious, but I am not sure it is convincing. Let us start by considering (SBP). I agree that, plausibly enough, (6) would still be true if its *present* supervenience base (if any) were wiped off. But I definitely *do not* believe that (6) would still be true if "*everything on which the truth of this proposition could plausibly be thought to supervene*" were wiped off. For it is very plausible to suppose that (6)'s truth supervenes on *the past* (on past events, say).[2] In my mind, this is the reason why the following counterfactual sounds implausible—or at least, much less plausible than (SBP):

(SBP*) If it is true that there was a sea-battle in 2019, it would still be true that there was a sea-battle in 2019, even if the past were annihilated, to the effect that any sea-battle-related event were erased from reality.

Some reader will be puzzled by the idea that *past events* can be erased. If so, try engaging in a suitable time travel fantasy, or assume that God can bring about changes in the past. But if you are a strict presentist like Todd, you need not do so. For in your mind, this is what always happens: past events (e.g., the moon landing, or Jack Kennedy's killing) are literally erased from reality as soon as they cease to be present and become past. Symmetrical considerations apply to (SBF). A counterfactual like (SBF) sounds implausible because we take for granted that the truth of future contingents supervene on the *future* (on future events, say), and that dramatic changes in present reality would causally affect such (future) supervenience base. This is the reason why the following, modified analogue of (SBF) sounds plausible:

(SBF*) If it is true that there will be a sea-battle in 2219, it would still be true that there will be a sea-battle in 2219, even if everything [on which the presentist might want to base the truth of this proposition] went out of existence, as long as this does not affect any sea-battle-related future event.

To clarify, let me express my disagreement with Todd's reasoning in a slightly different way. I agree with Todd that, based on our intuitions, we are inclined to accept (SBP) and reject (SBF). But I think that, before drawing any definite conclusion, one should first pause and ask *why* we have these intuitions. In my mind, (SBP) sounds plausible because we take factual truths about the past to supervene on *the past*, and given that the past is not causally dependent on the present, we are led to think that past-directed truths do not supervene on anything *present*. But this conclusion is different from the one Todd wants to draw, namely, that past-directed truths do not supervene on anything *at all*, even if I understand why a presentist may be tempted to equate these conclusions. In a nutshell, in my mind, (SBP) sounds plausible not because we take past-directed truths to be brute but because we implicitly assume that the past is part of reality. Parallel considerations hold for (SBF), taking into consideration that the present *does* causally affect the future, and so it *is* plausible to hold that future truths partly supervene on present reality. Be that as it may, I would like to point out that even if I find Todd's arguments about TSR unconvincing, this does not detract from the interest of Todd's main views, which are introduced and defended in the remainder of the book and are largely independent of the content of this chapter.

[2] See also Ingram 2023 for a similar complaint.

Chapter 2 (*Three Models of the Undetermined Future*) includes some of the most insightful parts of the book. Todd introduces three models of the future and makes it clear why his model of choice supports (OF). In describing these models, Todd presupposes the so-called *branching-time* (BT) conception of reality. In the BT conception, our tense-modal universe is represented as a tree-like structure of possible worlds (also known as *histories*). Worlds/histories are sequences of successive *moments*, where a moment represents a temporally instantaneous and spatially maximal event (a spatially complete temporal 'slice').[3] Any two worlds/histories in the structure *overlap* (i.e., share their moments with each other) up to a certain moment and then they *divide* or *branch*. Figure 1 depicts a very small BT universe, that only spans for three units of time, where *now* is the actual present moment.

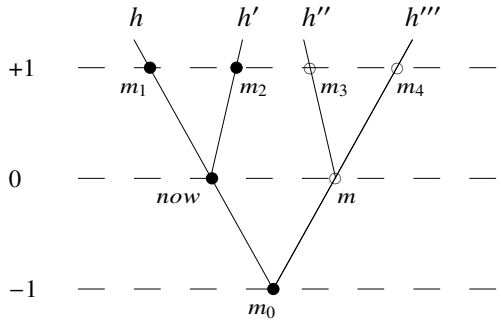


Figure 1: A tree-time-unit universe, where h–h''' are alternative histories, m_0 is the only past moment, *now* is the actual present moment, m is an alternative present moment, and m_1–m_4 are possible future moments. White circles represent moments that are *now* causally impossible.

Within the BT conception, causal indeterminism translates as the view that different histories are now *causally possible*, that is, their future stretch is a possible causal outcome of the present moment. The three models Todd discusses are all indeterministic in this sense, and they differ as to what *primitive future directed facts* (i.e., contingent facts about the future) are taken to be part of (present) reality.

Model (I) (Ockhamism) There is only one causally possible history consistent with the (primitive) future directed facts, and it is determinate which history this is.

Model (II) (Supervaluationism) There is only one causally possible history consistent with the future directed facts, but it is indeterminate which history this is.

Model (III) (Todd's view) There are many causally possible histories consistent with the future directed facts, because there are no future directed facts.

Todd complains that Model (I) and Model (II) are often conflated with one another, and I agree. I also perfectly agree with Todd that indeterminist presentists would do better to avoid Model (I). (As a matter of fact, Model (I) is, under some

[3] The existence of all-encompassing instantaneous events is incompatible with special relativity. The problem can be solved by resorting to branching *space*-time structures (see Belnap et al. 2022) but, arguably, this solution is not available to presentists.

description, the standard model of choice for B-theorists, just like Model (II) is for A-theorists.)

As for Model (II), it is based on a very natural thought. Suppose you flip an indeterministic coin. There are two possible future outcomes for your toss (heads or tails). However, intuitively, just one of them will be *actual*, even though it is now indeterminate which. The same holds for future outcomes in general: for at least some future time t (e.g., 12 PM CET 01.01.2043), several alternative moments are still possible but, by necessity, exactly one will obtain at t, although it is now indeterminate which.

The key idea behind Model (III) is that there are no (primitive) facts directed towards the future. (As we shall see, in the context of Todd's general framework, this idea results in a stronger tenet, namely, that there are no truths directed towards *later moments*.) Model (III) shares an important feature with (a B-theoretical understanding of) Model (I): both models allow for no ontic indeterminacy at the fundamental level. In Model (I), there is no indeterminacy as to what events are fundamentally (i.e., tenselessly) part of actuality (to be sure, the future is indeterminate *as of now*, but this is just a local, perspectival form of indeterminacy). Similarly, in Model (III), there is no ontic indeterminacy as to what facts hold (e.g., it is not indeterminate whether it is a fact today that there will be a sea-battle tomorrow; rather, it is not a fact that there will be a sea-battle tomorrow and it is not a fact that there will not be). The proponents of Models (I) and (III) agree that we can correctly describe an indeterministic universe without positing fundamentally indeterminate facts or events. (B-theorists hold the same view also toward *change*: they think we can correctly represent a changing universe without positing changes at the fundamental level.) Todd argues at length against attempts to show that Model (III) is ultimately incoherent. In my mind, he succeeds in this.

In Chapter 3 (*The Open Future, Classical Style*), Todd introduces and discusses his favorite semantics for future-tensed statements. In the philosophical debate on future tense, Todd's iconic view (OF) ("All future contingents are false") is generally associated with a semantic approach known as *Peirceanism*. Todd's overall semantic package agrees with Peirceanism on the truth-conditions of future tensed sentences, but there is an important difference between Todd's semantics and Peirceanism. Highlighting this difference helps make Todd's overall views clearer. Todd summarizes Peirceanism as the following semantic claim about future-tensed statements:

> (APF) It will be in n units of time that p iff in all of the causally possible futures, in n units of time, p (36).

Accordingly, a prediction like (1) turns out to be semantically equivalent to:

(1*) Russia will invade Ukraine in days as a matter of causal necessity.

This is very implausible. If you assert (1) based on intelligence reports, you do not mean that the invasion is causally inevitable. (No intelligence report can ensure you of *that*!) Todd's semantics is based instead on the following principle:

> (AAF) It will be in n units of time that p iff in all of the available futures, in n units of time, p (30),

where the futures that are *available* at a moment m are the causally possible futures consistent with the primitive future-directed facts that obtain at m, if any such fact exists. In Todd's mind, (AAF) provides a *neutral* semantics for sentences of form "It will be in n units of time that p" ($F_n p$ in symbols).[4] To get to the specific truth conditions of $F_n p$, we need to specify the underlying notion of *availability*, that is, take stance on what future-directed facts obtain. In other words, the truth-conditions we pre-theoretically assign to future contingents are not just a matter of semantics. They also depend on our metaphysics, namely, on the model of reality we pre-theoretically agree upon. According to Todd, we all take for granted that (there are future-directed facts to the effect that) there is exactly one available future: the actual future. In other words, we all presuppose that either Model (I) or Model (II) is correct. When this common presupposition is in place, (AAF) becomes equivalent to:

> (UAF) It will be in n units of time that p iff in the unique actual future, in n units of time, p (31).

In contrast, Todd adopts Model (III), according to which there are no future-directed facts. Therefore, availability collapses on causal possibility. But even if Todd agrees with Peirceans on the truth-conditions of a sentence of form $F_n p$, he denies that (AAF) is a purely semantical ('analytical') truth. Therefore, Todd agrees with Peirceans that (1) is equivalent to (1*) in the context of a certain metaphysical conception of the future (Model (III)), but he denies that they are semantically equivalent, that is, synonyms. So, Todd's view is not as implausible as Peirceanism.

To summarize, the main advantage of Todd's proposal over Peirceanism is that it comes equipped with a strategy to explain why (OF) and other key open futurist principles sound implausible. Let us call it the *metaphysical presupposition strategy*: many of the views Todd subscribes to sound implausible because we pre-theoretically presuppose that future-directed facts exist, to the effect that there is exactly one actual history. If, as Todd is prepared to do, we entirely give up this presupposition ("The actual world is an ontological or metaphysical posit that can be dispensed with" [78]), then Todd's view becomes plausible or at least, less crazy than it might appear at first sight.

In the book, Todd uses the metaphysical presupposition strategy to account for other counterintuitive consequences of (APF). Here are two examples.

Scopelessness. *Will* is apparently scopeless with respect to negation (and arguably, other operators as well): there is no perceived difference in meaning between sentences of the forms $F_n\neg p$ and $\neg F_n p$. But (APF) entails that there is a deep semantic difference between these forms (e.g., future contingent statements of form $F_n\neg p$ are all false, while the corresponding statements of form $\neg F_n p$ are all true). Thus, (APF) is empirically inadequate.

Wrong propositions. If you fear, based on evidence, that tomorrow you will be tortured, and you follow Todd in believing that all future contingent propositions are false, then your fear should be irrational. But this does not sound

[4] As a matter of fact, I do not think that (AAF) is neutral. At most, it may sound neutral if you ignore B-theorists, who do not need to resort to future-directed facts (neither in general nor) in specifying the truth conditions of future-tensed claims.

> quite right. Similar problems can be restated with reference to other propositional attitudes: it appears that, by adopting (APF), we end up systematically pairing off future contingent that-clauses with the wrong propositions.

It should be clear how the metaphysical presupposition strategy works in these cases. As for *Scopelessness*, Todd replies that we regard $F_n\neg p$ and $\neg F_n p$ as semantically equivalent because, in assessing them, we presuppose that there is only one available future—and indeed, $F_n\neg p$ and $\neg F_n p$ are equivalent under this presupposition. And here is Todd's reply to *Wrong propositions*:

> I am inclined to say that the problem here simply reveals how deeply our bias is towards the view that there is a unique actual world. If there is a unique actual world, then when we learn that it is false that we will be tortured tomorrow, we learn that we won't be tortured tomorrow—and so there is nothing left to fear. But if there just is no such thing as the unique actual world, then even if we learn that it is false that we will be tortured tomorrow, that doesn't tell us that we won't be tortured tomorrow—we certainly could end up being tortured tomorrow, and so there is indeed something left to fear. I suppose this amounts to me simply biting [the] bullet (95).

Thus far, I have only focused on (UAF) and (APF). But these are not the only semantics in town. A key alternative to (AAF) are proposals based on *postsemantic* clauses, such as Thomason's (1970) *supervaluationism* and MacFarlane's (2003) *relativism*. In the chapter, Todd compares his proposal with supervaluationism, but virtually all he says extends to other postsemantic proposals. The key advantage Todd claims for his proposal over postsemantic proposals is its full consistency with classical logic.

> **Fully classical.** (APF) results in a temporal logic that is fully classical, as it posits no truth-value gap and treats Boolean connectives as truth-functional operators.

What I said above about Todd's metaphysics also holds for Todd's semantics: (APF) is an A-theoretical proposal that shares many features (and benefits) of B-theoretical approaches. And I agree with Todd that there is something puzzling about supervaluationism when it comes to the behaviour of connectives—e.g., it is not clear how a disjunction can possibly be true when neither of its disjuncts are.

From a metaphysical viewpoint, Todd contends that supervaluationism, being based on Model (II), crucially differs from his own proposal in that it assumes that a unique actual future exists ("[t]he crucial posit of the supervaluationist is thus that *there is an actual world*" [77]). As I have argued elsewhere (Spolaore and Gallina 2020), I agree with Todd that supervaluationism *is consistent* with the assumption of a unique actual future, but I do not think it requires that assumption, and the same holds for other postsemantic proposals. Thus, I do not think that the key diference between (APF) and postsemantic proposals concerns the existence of the actual future. (And neither it concerns (OF), for there are postsemantic proposals in which (OF) is true; see Iacona and Iaquinto 2023.) Rather, to employ Todd's notion of *availability*, I think that the key difference between Todd's view and postsemantic proposals lies in the *semantic role* they assign to availability. Let me explain, focusing on supervaluationism for simplicity.

Supervaluationism and (APF) agree in equating availability with causal possibility. The crucial difference between them is that, in supervaluationism,

availability is fixed by the context, while in (APF) it is moment-relative, and so it is shiftable by tense operators. To illustrate, consider again Figure 1, and suppose that the present-tensed sentence p is true at the present moment *now*, and false at any other moment. Moreover, suppose that *now* you utter a sentence of form:

(7) $P_1F_1\,p$

It is easy to check that this utterance comes out true or false depending on whether availability is understood as context-dependent or moment-relative. If availability is context-dependent, as in supervaluationism, then the available moments are all and only those that are causally possible *relative to the moment of the context* (i.e., *now*). Thus, in the model depicted in Figure 1, the only histories (contextually) available are h and h'. Given this context-dependent notion of availability, (7) comes out true, for 1 unit of time before *now* (at m), on all available future (on both h and h'), in 1 unit of time (*now*), p. If availability is moment-relative, as Todd assumes, then (7) comes out false as uttered *now*, for 1 unit of time before *now*, at m, on some histories that *were* causally possible at m (namely, on h'' and h'''), in 1 unit of time (at m), it is not the case that p.

Todd's view presupposes a moment-relative, shiftable notion of availability, while postsemantic proposals presuppose a context-dependent notion of availability. This is, in my mind, the key difference between them and, it seems to me, it is semantic and not metaphysical in nature (see also MacFarlane forthcoming for a similar point). This difference has immediate impact on the retrospective evaluation of future contingents. For instance, supervaluationism validates the following principle:

> **Retrospection.** If p is true in context c relative to the moment of the context, then $F_n\,p$ is true in context c relative to a moment m (if any) that precedes by n units of time the moment of the context.

For instance, if we know that today there is a sea-battle, then today we are in a position to assess "Tomorrow there will be a sea-battle" as true relative to yesterday.

Let us note that Todd cannot reply that, if one regards "Tomorrow there will be a sea-battle" as true relative to yesterday, one is committed to the metaphysical view that yesterday it 'was already a fact' that there would be a sea-battle in one day. For this consequence only follows if we assume (APF). In supervaluationism, the truth of "Tomorrow there will be a sea-battle" relative to yesterday requires only that *at the moment of the context* (viz., *today*) it is a fact (or a truth) that, yesterday, there would be a sea-battle in one day. For similar reasons, even if supervaluationists subscribe to the *retroclosure principle* ($p \rightarrow P_n F_n\,p$), this does not mean that they are committed to the view that, n units of time ago, some (then) future-directed fact 'already obtained'. And the same goes, for similar reasons, for the principle Todd calls *will excluded middle* (WEM, $F_n\,p \vee F_n\neg p$)—at least to the extent that supervaluationists endorse it.[5]

[5] Actually, as Todd recognizes, WEM is not supervaluationistically valid. A valid principle in the surroundings is $F_n(p \vee \neg p) \rightarrow (F_n\,p \vee F_n\neg p)$, which says that WEM is valid provided either p or $\neg p$ will be true in n units of time. The key supervaluationist argument for this principle runs as follows. Let us focus on Figure 1 and on the instance $F_1(p \vee \neg p) \rightarrow (F_1 p \vee F_1\neg p)$ of the principle, supposing it is uttered *now*. Thanks to the antecedent, we know that in 1 time unit, a moment m^* will be present, in which either p or $\neg p$ is true. Now *let us pretend that* 1 *unit of time have elapsed*, and m^* is present (is the time of the context). Thus,

What I think Todd can contend is that, if no-futurists adopt (APF) over postsemantic proposals, they get a language that is *more transparent* with respect to their underlying metaphysical commitments. For at each step in the recursive evaluation of a sentence, (APF) requires to evaluate as true at a moment m only those (sub-)sentences that are based on facts (or truths) that obtain at m. In contrast, supervaluationism allows a certain degree of metaphysical irresponsibility: whether a statement is true relative to a moment or in a certain context does not strictly depend on facts that obtain at that moment/context. My personal impression is that supervalutionism is much more plausible as a theory concerning the *assertibility* conditions of future-tensed statements than as a theory concerning their truth conditions, especially if we are interested in the metaphysical conditions grounding their truth. Be that as it may, I think that Todd could argue that his semantics is best suited for *philosophical discussions* within the no-futurist camp, even at the price of some clashes with ordinary semantic intuitions. If I am right, this is a real advantage of Todd's proposal, which he alludes to somewhat in the book, but without emphasizing it much.

In Chapter 4 (*The Would-Will Connection*), Todd highlights a few semantic analogies between future contingents, as he understands them, and counterfactual conditionals. Todd assumes a standard, Lewis-Stalnaker semantics for counterfactuals conditional, in which a counterfactual of the form "If it were the case that p, then it would be the case that q" ($p \Rightarrow q$ for short) is true when q is true in all antecedent-worlds[6] most similar to the actual world. According to Todd, WEM ($F_n p \vee F_n \neg p$) is analogous to the so-called *conditional excluded middle* (CEM, ($p \Rightarrow q$) $\vee$ ($p \Rightarrow \neg q$)). The analogy ultimately consists in that both principles call into question a view Todd calls *the grounding claim*: future contingents and counterfactuals must both be grounded in actuality. According to Todd, accepting WEM as valid (as both the advocates of Ockhamism and of postsemantic proposals do) amounts to denying that future-tensed truths just depend on how actual (i.e., present) things are,[7] just like accepting CEM as valid amounts to denying that counterfactual truths just depend on how actual things are like. The analogy also helps Todd deal with *Scopelessness*. Many philosophers reject CEM as invalid and, as a consequence, they deny that *would* is scopeless as it occurs in counterfactuals. Given the analogy between *will* and *would*, Todd argues, these philosophers should at least be open to the view that *will* is not scopeless, either.

Chapters 6 (*Betting on the Open Future*) mostly deals with a problem for open futurists that we can call the *bet problem*.[8] Consider the bet that a certain horse, Phar Lap, will win a certain race.

either p or $\neg p$. If p, then (by retrospection) $F_1 p$ is true relative to *now* in context m^*, otherwise $F_1 \neg p$ is. Either way, we are sure that $F_n p \vee F_n \neg p$ is true relative to *now*, no matter what moment will be present in 1 time unit. Thus, we can assert it in the actual present context (*now*), even though neither the truth of $F_1 p$ nor the truth of $F_1 \neg p$ is grounded in present facts.

[6] An antecedent-world is a world where the antecedent of the conditional (here, p) is true.

[7] As mentioned above, I am not much in agreement with Todd that, to the extent that they subscribe to WEM, supervaluationists are committed towards future-directed facts.

[8] The chapter also discusses how open futurist should deal with probabilistic talk about the future—a topic that I cannot cover here due to space limitations.

> [The problem is that, since] on the open future view, the proposition "Phar Lap will win the race" was not true at the time of the given bet, it follows that anyone who had bet that Phar Lap would win the race would fail to win the bet—even from the perspective of a time at which Phar Lap has in fact won (120).

In other words, it appears that open futurists have a problem in making sense of our practice of betting, for it makes little sense to bet on the truth of a future contingent proposition if all such propositions are bound to be false. This sounds like an instance of the *Wrong propositions* problem, and Todd *could* address it by resorting to the metaphysical presupposition strategy. But he believes that a stronger reply is possible in this specific case.

In Todd's mind, there is a hidden assumption behind the bet problem, namely, that a bet that *p* is a bet on the truth of the proposition that *p*. And Todd's approach to the problem consists precisely in denying this assumption. According to Todd, the practice of betting has to do with our conditional behavioural commitments, and not (necessarily) with propositional truth.

> My suggestion is that when one bets on Phar Lap to win, one is not betting (or at least *need not* be betting) on the current truth of the proposition "Phar Lap will win". Rather, one is directly *bringing it about* that any future in which Phar Lap wins the race is *thereby* a future in which one wins the bet [...]. In other words: One *bets* that it will be that *p* iff one does something that brings it about that any *p* future is a future in which one is owed the [contextually specified] betting response (121).

In the end, according to Todd, "when it comes to betting on future events which have not yet transpired, *current truth* is irrelevant" (121).

I shall limit myself to discussing one of the arguments Todd provides for this conclusion. Todd asks us to consider the following dialogue, where both A and B are open futurists.

> A: It is not now true that there will be rain tomorrow, and not now true that there will be no rain tomorrow. It is open. But let's agree: if it does rain tomorrow, you owe me £5, and if it does not rain tomorrow, I owe you £5. Agreed?
> B: Agreed (124).

Clearly, A's discourse is perfectly felicitous. But, Todd argues, it should sound confusing (if not outright incoherent), assuming A is betting on whether the proposition that it will rain tomorrow *is* true.

As I said towards the beginning of this paper, I think that time-relative truth ascriptions are ambiguous. When we say that a proposition *is now* (or *currently*) true, we often implicate that the proposition is *historically necessary*. Thus, I agree with Todd that we generally do not bet on the *current truth* of propositions in this sense, because current truth in this sense just means historical necessity. As for Todd's dialogue, my impression is that A's discourse sounds felicitous for essentially the same reason that my above example (5) does: because to bet/assert that the proposition *p* is *now* true—in at least one important reading of "now true"—is not equivalent to bet/assert that *p*.

The eighth and last chapter (*The Assertion Problem*) deals with the so-called *assertion problem* for open futurists, which is really a bunch of different problems.

Here I shall limit myself to discussing two of them. First, there is a problem close to *Wrong propositions*: if future contingents are all false, why do we assert them all the time? Todd's basic response consists, again, in the metaphysical presupposition strategy: people happily assert predictions about future contingent events because they presuppose the existence of a unique actual future. The second problem has to do with the open futurists' own behavior. Open futurists happily assert and accept future contingents all the time. But, so one could object, they should not, given that they supposedly *know* they are false. They should refrain from asserting them, and possibly even correct people who assert them. Todd eloquently argues against these normative conclusions by appeal to the analogy between (OF) and Trenton Merricks's eliminativism towards ordinary objects.

> [T]he eliminativist believes that there are no tables and chairs—really, there are only atoms arranged tablewise, and atoms arranged chairwise, and so on. Now consider. What follows concerning what the eliminativist should and should not assert, and what follows concerning what assertions (from others) the eliminativist should and should not attempt to correct? On latter issue first [...]. It is perfectly obvious that Merricks is under no obligation, given his acceptance of eliminativism, to attempt to correct any such ordinary person speaking in the midst of life. To suggest that this [is] what Merricks *should* do, given that he accepts the relevant theory, is to suggest that Merricks should, *inter alia*, consistently waste his own time, and annoy and confuse a host of innocent bystanders in the process. [...] Thus: even if he believes eliminativism, and even if eliminativism is true, it is not the case that Merricks *should* (attempt to) correct people in the imagined way in *ordinary life*. And, I suggest, the same is true of himself: it is not the case that Merricks should (attempt to) monitor *himself* in ordinary life, making sure, for instance, always to talk in terms of atoms arranged table wise, and never in terms of tables. That would be a terrific waste of mental energy, and more else besides. It just doesn't matter (185-86).

For similar reasons, Todd concludes, it would be absurd to require open futurists to change their ordinary way of speaking and behaving because of their arcane metaphysical views.

Finally, Todd contends that future contingents, even if all false, can serve valuable communicative purposes. The reason is basically that, in many cases, when we make a future-tensed claim, in addition to asserting a (false) proposition about the future, we convey or suggest valuable information about plans, intentions, and tendencies. Here is an example.

> "Look. It will rain all weekend".
> Falsehood asserted: It will rain all weekend.
> Truth conveyed/suggested replacement talk: The world is tending toward rain all weekend.
> If things don't change, it will rain all weekend (195).

The book concludes with the following words.

> I hope we can now take seriously—or more seriously—the position that future contingents are systematically false (202).

And I am happy to grant that the hope is well grounded: after Todd's book, we should definitely take (OF) more seriously.

Before wrapping up, let me point out two key problems of Todd's view, which in my mind he does not adequately discuss in the book. First, Todd tends to deflect the contention that his views are counterintuitive by appeal to a *good company* argument: all kinds of philosophers have held counterintuitive views. As the most prominent example, Todd mentions Merricks's eliminativism about ordinary objects. I think, however, that there are key dialectical differences between Todd's open futurism and Merricks' eliminativism. We now *know*, based on strong scientific evidence, that our commonsense picture of ordinary objects as solid, well-separated sources of causal influence is false. Thus, some piece of that picture must be given up, and Merricks's eliminativism is just an especially radical way of revising the picture. But there is no comparable scientific pressure on our common understanding of future contingents. The pressure only comes from *presentism* (or *no futurism*), a view entirely based on metaphysical principles and reasonings—i.e., on refined common sense. Thus, there is something self-defeating in Todd's proposal which is not in Merricks's. We already knew that presentism is (at best) at odds with scientific theories and practice, and we now know (thanks to Todd) that, if worked out in a perfectly coherent and "classical" fashion, it has severely counter-intuitive consequences. In light of this, it sounds like the best thing to do is to give up presentism, does not it?

The second key problem has to do with the *Wrong propositions* problem and its relationships with the BT conception. Consider again (APF):

(APF) $F_n p$ iff in all of the causally possible futures, in n units of time, p.

Let us abbreviate the phrase "in n units of time, p" as $f_n p$ (with small f). Clearly, (APF) presupposes that, at any moment m, a sentence of form $f_n p$ has a specific truth-value relative to any history passing through m. But this means that each instance of $f_n p$ expresses an (unstructured) proposition, that is, the class of worlds where it is true. Todd cannot deny this, for otherwise his entire proposal ((APF), and also (AAF)) would be unintelligible. It is very plausible to suppose that $f_n p$ translates as a future-tensed sentence in English—just like $F_n p$. To avoid ambiguities, let us use an italicized "*will*" in translating $f_n p$. Now consider an example of the *Wrong propositions* problem. At a moment m, you sincerely assert:

(8) I fear that there will be a sea-battle tomorrow.

Clearly, your fear is not directed toward the (necessarily false) proposition that, according to (APF), the sentence of form $F_n p$ "there will be a sea-battle tomorrow" expresses at m. But what about the proposition expressed at m by the sentence of form $f_n p$ "there *will* be a sea-battle tomorrow"? This proposition corresponds to the class of histories compatible with m where a sea-battle obtains tomorrow, it is a contingent proposition, and it fits perfectly well the role of your object of fear. The obvious conclusion is that, in (8), "there will be a sea-battle tomorrow" is to be translated as a statement of form $f_n p$ ("there *will* be a sea-battle tomorrow") and not as a statement of form $F_n p$. I cannot see how Todd can possibly reject this obvious conclusion, for he cannot deny that statements of form $f_n p$ express perfectly intelligible propositions, and there is no reason to suppose they have no English counterpart. But for some reason, Todd is clearly not prepared to accept it. We can call this problem *Right propositions*: if (APF) is

intelligible and meaningful at all, then history-relative future-tense claims of form $f_n p$ must express propositions, and these are precisely the contingent propositions speakers intend to express by uttering future contingents, at least when they occur within propositional attitude ascriptions; but then, uncontroversial interpretative principles force us to translate those utterances as sentences of form $f_n p$. But this is a conclusion that Todd, rightly or wrongly, is clearly not prepared to accept.

Let me conclude this critical discussion with some general remarks on the value of Todd's book. As I said at the beginning, despite my general philosophical inclinations being very different from Todd's, I really enjoyed *The Open Future*. It is an excellent book. It is engaging, thought provoking, and the breadth of topics it covers is impressive. Todd's arguments and observations are mostly precise, subtle, and original. The views he defends are very counterintuitive at first sight but are surprisingly resilient to refutation. Moreover, and more importantly, I think that presentists should definitely take Todd's semantic proposals seriously, especially if they are interested in adopting a logic whose validities closely mirror their metaphysical assumptions. Which doesn't hurt, the book is excellently written, with a charming and lively style. I wholeheartedly recommend it to all philosophers interested in the interplay between metaphysics and philosophy of logic, and it is obviously a must-read for anyone working on future contingents, divine foreknowledge, and the semantics of future tense.

References

Belnap, N., Müller, T., and Placek, T. 2022, *Branching Space-Times: Theory and Applications*, Oxford: Oxford University Press.

Iacona, A. and Iaquinto, S. 2023, "Postsemantic Peirceanism", *American Philosophical Quarterly*, 60, 3, 249-56.

Ingram, D. 2023, "Review of Patrick Todd, *The Open Future*", *Metaphilosophy*, 54, 364-67.

MacFarlane, J. 2003, "Future Contingents and Relative Truth", *Philosophical Quarterly*, 53, 212, 321-36.

MacFarlane, J. forthcoming, "Why Future Contingents Are Not All False", *Analytic Philosophy*.

Spolaore, G. and Del Prete, F. 2019, "Now There Will Be Trouble", in Hasle, P., Blackburn, P., and Øhrstrøm, P. (eds.), *Logic and Philosophy of Time: Further Themes from Prior*, Aalborg: Aalborg University Press.

Spolaore, G. and Gallina, F. 2020, "The Actual Future Is Open", *Erkenntnis*, 85, 1, 99-119.

Thomason, R.H. 1970, "Indeterminist Time and Truth-value Gaps", *Theoria*, 36, 3, 264-81.

Tooley, M. 1997, *Time, Tense, and Causation*, Oxford: Oxford University Press.

The Affective and Practical Consequences of Presentism and Eternalism

Mauro Dorato

Rome 3 University

Abstract

In the dispute between presentism and eternalism, the affective dimensions of the debate have been somewhat neglected. Contemporary philosophers of time have not tried to relate these ontological positions with two of the most discussed maxims in the history of ethics—"live in the present" vs. "look at your life under the aspect of the eternity" (*sub specie aeternitatis*)—that since the Hellenistic times have been regarded as strictly connected with them. Consequently, I raise the question of whether the endorsement of one of these two ontological views can make a practical difference in the way we should live.

Keywords: Presentism, Eternalism, Eternalism's affective attitudes, Presentism's affective attitudes.

1. Introduction

Despite the recent flurry of papers dealing with the relationship between presentism and eternalism and our temporally oriented attitudes, the *affective* dimensions of the debate have been somewhat neglected.[1] To clarify my main aim, which is to study and make explicit these dimensions, it is important to state at the outset what the presupposition and focus of my paper are. *First*, I will assume without further arguments that the debate between presentism and eternalism, however these views are formulated, is genuine and non-trivial (see premise (i) in Section 2 below), for instance because it is based on an unrestrict-

[1] Silverstein 1980, Bradley 2004, Le Poidevin 1995, Burley 2008, Finocchiaro and Sullivan 2016 have concentrated on the relationship between eternalism and the fear of death. The relation between eternalism and the notion of midlife crisis is at the center of a paper by Setiya 2014. Dorato 2008 and more recently Ismael 2016 and Deng 2017, have discussed a possible ethical reading of presentism. Greene and Sullivan 2015, Finocchiaro and Sullivan 2016 argue that we should treat all moments of our life on a par. Orilia 2016, who discusses presentism and eternalism from what he calls an existentialist viewpoint, defends the "open future" view arguing that we are free to shape, at least in part, the future.

Argumenta 9, 1 (2023): 173-190
ISSN 2465-2334

© 2021 Mauro Dorato
DOI 10.14275/2465-2334/20210.dor

ed view of existence.[2] *Second*, for my purpose I can safely ignore discussing other temporal ontologies (the growing block view of time),[3] and "mixed views", in which it is postulated a now moving on a tenselessly conceived series of events (Skow 2015). *Third*, I will also ignore the important issue of temporal neutrality, that is, the presence of time biases in the context of *rational decision theory* (see among many others, Brink 2011; Green and Sullivan 2015; Finocchiaro and Sullivan 2016; Callender 2017, chapter 12; Sullivan 2018). *Fourth*, I will assume that the reasons to believe in one of the two ontological views are independent of their affective consequences,[4] even if it were psychologically possible to end up believing in one ontological position because we would be better off if it were true. In other words, the fourth assumption—not be to be discussed here—is that our beliefs in one of the two ontologies is and ought to be influenced only by epistemic arguments.

There are at least two reasons to direct our attention to the issue whether a belief in one of the two opposite ontologies can have affective consequences that can make a practical difference in our lives.[5] The first is *historical*: not many contemporary philosophers of time have tried to relate presentism and eternalism, regarded as *ontological* positions, with two of the most discussed practical and affective maxims in the history of ethics namely—"live in the present" or "seize the day" (affective presentism)—vs. the Spinozistic maxim "look at your life *sub specie aeternitatis*"[6] (affective eternalism). And yet, since the Hellenistic age, philosophers *motivated* the adoption of (i) pragmatic presentism (how we should act) with the ontological claim that only the present exists and (ii) the adoption of affective eternalism with ontological eternalism.[7] As in assumption 4 above, both these claims were backed up by purely epistemic arguments. The second motivation is more *theoretical*: to name one, Deng (2017) has argued that the dispute between eternalism and presentism reduces, as a matter of *meaning*, to the problem of which of the *two emotional or affective attitudes* toward the two ontic positions is preferable.

[2] With apologies to many philosophers that here I do not mention, the lack of cognitive content of the ontic debate has been defended by Meyer 2005, Savitt 2006, Dolev 2007, and Deng 2017. With the same proviso, philosophers supporting the contrary view are Crisp 2004, Merricks 2007, Mozersky 2011, and Sider 2016.

[3] For a presentation of the growing block theory, see, among others Pooley 2013, and more recently, Correia and Rosenkranz 2018, which, to my knowledge, is the best book-length defense of the growing block theory of time against the objections to be found in the literature.

[4] On this point, see Mellor's 1999 criticism of Cockburn 1997.

[5] The term 'affective' will be clarified in the following.

[6] A translation of this Latin expression could be "from an eternal point of view", or "under the aspect of eternity".

[7] Ismael refers to this attitude as the "temporally transcendent view" of our life (see 2016: 226) and contrasts it with the caught-in-the-moment view. Here I try to discuss these issues in more detail and consider practical eternalism as the view that our life is an inseparable part of the cosmic order.

2. The Argumentative Structure and the Plan of the Paper

In order to evaluate Deng's important conclusion, I will (paradoxically) start with a contrary premise, which, as anticipated above, for the sake of the argument will be taken for granted:

(i) (Unlike Deng's claim), the ontological dispute is not merely verbal.
(ii) There are two different affective attitudes related, respectively, to ontic presentism and eternalism that are influenced only by our purely epistemically motivated beliefs in the two respective ontologies: *namely affective presentism and affective eternalism as defined before*
(iii) Our *actions* are at least partly motivated by our affections or emotions.

Conclusion: these different attitudes toward time make some *practical difference* in how we act (and possibly should act).

In a word the conclusion argues that a believer in ontological presentism will feel, act and should act differently from a believer in ontology presentism. For lack of better terms, I have labelled *affective presentism* and *affective eternalism the emotional attitudes toward time that, if I am right, can be respectively attributed to believers in the respective ontological positions*. A more precise account of the two affective attitudes referred to by the maxims above will be given in the following sections.

Let us now discuss the premises of the argument. Premise (i) was *assumed* to be true. Premise (ii) is the object of the following discussion. Premise (iii) is based on a lot of empirical evidence to the effect that emotions motivate our actions. The conclusion of the argument would be very interesting if the premises were reasonable: an epistemic commitment to the two ontological views makes some practical difference in the decisions and the actions of presentists and eternalists. An important objection to the conclusion of this argument will be discussed in the last section of the paper.

More in detail, the plan of the paper is as follows. In the next section, by taking advantage of some quotations from the history of ethics, I will illustrate the sense in which, at least since the Hellenistic age, the two ontological stances about time have historically been regarded as one of the main instruments to live a flourishing life[8]. It then becomes important to establish whether these historical arguments can be justified and relied upon also today. The fourth and the fifth sections will clarify as precisely as possible, respectively, the meaning of the metaphoric expressions "live in the present" and "look at your life sub specie aeternitatis", which are used in the previous sections in an intuitive way. It should be obvious why, without such a clarification, there cannot be any precise discussion of the link between affective and practical attitudes toward time and the respective ontological beliefs. In the last section, I will discuss the philosophical consequences for the ontological debate by evaluating an important objection to the conclusion of the argument above, which consists in claiming that the two temporal attitudes can be experienced by the *same individual in different moments of her life*, so that an ontic presentist can be a practical eternalist and vice versa.

[8] For a reconstruction of the Hellenistic philosophy, see Inwood and Gerson 1997.

3. Following the Past Philosophers' Call to Live in the Present

The historian of ancient philosophy Pierre Hadot has well documented how in all Greek philosophical schools the meditation on the finiteness of our life and therefore on death was regarded as the essential step to learn to appreciate the inestimable value of the present moment (Hadot 1995: 28). We can add that in a logically possible world in which we knew that we could live for an extremely long amount of time by experiencing also courses of actions that we have given up today, there would be no necessity of choosing to do what is most important for us here and now by neglecting trivialities: the possibility of forming a character and a personality would be lost.

It should be clear why the link between dynamical presentism, a mind-independently conceived passage of time, consisting in the coming to be of previously nonexistent future events, and death is very robust. One of the strongest arguments in favor of the mind-independence of passage is our awareness of having to die in a more or less distant future. The passage of time, as dynamical presentism has it, implies that, independently of all problems raised by a *literally* moving now,[9] the commonsensical claim that *each day we are one day older* is non-tautological but simply undeniably true. In more respectable words, the claim that time passes simply *means* that, relative to today, each passing day the interval of time separating our present experience from the moment of our birth is one day *longer* and that, correspondingly, whenever our death will come, we know that each passing day the interval of time separating the present experience from our last day is one day *shorter* as in a burning fuse (Norton 2014).[10] To be more dramatic and avoid possible charges of selling tautologies of the kind "time passes one second per second", we could note that, by counting the passage of time in terms of our heart pulses, the number of heart pulses grows on average each minute by approximately 70 units, until our heart will stop.

The eternalist may object to these arguments in favor of a "dynamic" time in various ways. One is to paraphrase the statements above by introducing a finite number of unchanging B relations (to the extent that *days* are involved, of course). One unchanging relation links the day in which I am writing this paper to the day of my birth, the other linking today with the day of my death. And so on for each day until I die. In this God's eye point of view, which is not at all absurd, it is obvious that any dynamic element is lost: all of these relations don't change. However, for our purposes it will be sufficient to stress that the eternalist has to explain why we seem to find ourselves in different regions of spacetime in a purely relational way, as well as the fact that we first anticipate the same event, then we experience it and then we remember it.[11]

[9] For acute arguments against the common objections to "the one second per second" argument see Maudlin 2007. For a thorough defense of the moving now conception of becoming see Skow 2015.

[10] On a shrinking view of the future, see also Casati and Torrengo 2011.

[11] This problem has been recently voiced by Weatherall in his review of Callender 2017: "But what I still do not understand is how I, or anything else, get from one location or region of my worldline to any other. In other words, it is not merely that I represent myself to myself as occupying successively different locations in spacetime, with different stimuli, etc. It is also that, wherever I happen to be in spacetime, I will presently be elsewhere, and then elsewhere, inexorably. How does that happen?" Weatherall 2020: 6-7.

If for the sake of the argument we endorse this dynamical picture of time, we are ready to discuss the main affective and practical consequences of presentism. Even though *both* presentists and eternalists are aware that the duration of our life is limited, and that we live in a state of constant uncertainty, only dynamical presentists in the sense specified above can literally make sense of the claim that the time of our death *nears* one day every day. In a characteristic Epicurean spirit, Horace wrote:

> Inquire not […] how long a term of life the gods have granted to you or to me: neither consult the Chaldean calculations. […] Whether Jupiter has granted us more winters, or [this is] the last […]. Be wise; rack off your wines, and abridge your hopes [in proportion] to the shortness of your life. While we are conversing, envious age has been flying; seize the present day, not giving the least credit to the succeeding one (Horace, *Book 1*, *Ode 11*).[12]

The awareness, obviously shared by the eternalists, that our life has but a finite temporal extension and could soon end is a powerful drive to avoid (as much as it is reasonable) hopes and fears generated by imagined future events, and focus on the present experience. However, the question is whether the presentist's belief in a dynamical passage of time—that is not just *felt* like a subjective *quale* but refers to what she takes to be an objective metaphysical fact—can be very effective in changing her affective stance, emotions, and therefore practical decisions that lead to live her a more flourishing life.

For instance, the affective role of death in practical presentism is illustrated in a very clear way by a letter that Epicurus sent to Menoecus, in which he claims that worrying about our death *in the present* is irrational, because as long as we live, our death *does not exist* (we can add "in an unrestricted sense" as in contemporary literature). Consequently, painful anticipation in the present of an event that in principle we cannot directly experience now should play no role in our mental life:

> Foolish, therefore, is the man who says that he fears death, not because it will pain when it comes, but because it pains in the prospect. *Whatever causes no annoyance when it is present, causes only a groundless pain in the expectation*. Death, therefore, the most awful of evils, is nothing to us, seeing that, when we are, death has not come, and, when death has occurred, we are not (my emphasis).[13]

Our death does not strictly speaking belong to our life, it is only the process of dying that does. But even the process of dying ought to be "nothing for us", if this process is not experienced in the present. The reason why "Whatever causes no annoyance when it is present, causes only a groundless pain in the expectation" is given simply by the fact that the process of dying is a *future* event that relative to now does *not* exist. In sum, without doing too much violence to the strict meaning of the above letter, we can interpret Epicurus as claiming that it is rational (if at all) to suffer only for events occurring in the present: if past and future events don't exist, they are merely ghosts imagined by our minds in memory or "in the expectation".

[12] Here I am using the translation by Harrison 1981.
[13] Inwood & Gerson 1994: 28. For a contemporary debate on the role of eternalism in liberating us from the fear of dying, see note 1.

Note that Epicurus does not claim that we should accept ontic presentism for its practical advantages. On the contrary, it is from the indubitable belief that our death does not exist in the present that is irrational to fear it in the present. However, the irrationality of the felt belief can only be acquired after a considerable amount of "mental exercises" in the sense of Hadot (1995).[14] In this example, according to Epicurus, mental reflections on presentism as an ontological doctrine has the consequence of changing our emotional attitudes by changing our belief that being deprived of any capacity of experiencing the world is bad. The presentist can play the same argument not just for the remembrance of past traumatic events that exist only in our *present* thought, but also for those past events that precede our birth or non-existence that do not belong to our life. Since we don't fear the former, we should not fear the nothingness of the after-death state.[15]

A deep concentration in our present experience can even take us "outside of time" altogether. Wittgenstein echoed Epicurus' stress on the importance of living in the present as a way to escape the finitude of our existence, to which death does not belong: "Death is not an event of life. Death is not lived through. If by eternity is understood not endless temporal duration but timelessness, then he lives eternally who lives in the present" (Wittgenstein 1961: 6.4311).

In a word, the present moment, if lived fully and not halfheartedly, may even transmute into something eternal and take us outside time as in a mystical experience. The sense of timelessness pressed by Wittgenstein in the famous passage above is that eternity does not mean eternal duration, but a *nunc stans* (a standing now),[16] in which a present, intense experience annihilates our *experienced passage of time* that usually comprises memories and anticipations.

As evident from these quotations, the key reason that justifies focusing on the present moment is given by the fact that, as a consequence of presentism, there cannot be real happiness except now, because neither the past nor the future exist.[17] However, since all of our experiences occur in the present and it is only in the present that we can have happy (or unhappy) memories or happy (or unhappy) anticipations, where is the difference between the affective timbre of the presentism and eternalism?

[14] The sense of this word is more or less "constant mental practice" (on which we will not enter) whose purpose is to train our beliefs to become more and more adequate to the ontological conviction.

[15] The same mental effort should be cultivated to overcome our fear for the existence of another world: a rational understanding of the irrationality of our anxious anticipation of something that for us does not exist in our present experience serves the purpose of making the most of our present experience without groundless fears. There is huge literature on the "Symmetry Argument and Lucretius Against the Fear of Death". This is a title of a paper by Rosenbaum 1989. Lucretius claimed that if we look back at the eternity that passed before we were born, and mark how utterly it counts to us as nothing, we may see as in a mirror the time that shall be after we are dead.

[16] The term "nunc stans" is found in Boetius (475-526 A.D.) and was then revived by the 12th century philosophers: "The passing now makes time, the standing now makes eternity"; here 'standing now' means eternity as a property of God, who is outside time altogether.

[17] This viewpoint had been already defended by Aristippus of the Cyrenaic school in the fifth century B.C. For the notion of happiness in the ancient world, see Annas 1995.

The answer is that since the eternalist—as we will see in more detail in the subsequent section—recommends a self-transcendent view of time (see note 7), focusing her emotions only on the *present* internal and external experience would be to some extent *irrational*, given that for her the now exists on a par with the present and the future. This essentially means that for the eternalist the intentional past or future *content* of the *present* mental acts or states is (and should be) at least *sometimes* if not often be directed to non-present events, where the 'sometimes' is a proviso making room for the needs of everyday life.

The presentist's belief that the presently remembered events do not longer exist (unrestrictedly) may have (or ought to have) as a consequence focusing her emotions only on external events happening in the present, with the addition of those mental acts whose content *excludes* events occurred in the past (or anticipated in the future). The capacity for concentrating on the present experience in the sense explained above can de facto become more effective by realizing that—as a coherent consequence of an epistemic endorsement of *ontic* presentism—dwelling on the *memory* of an event that does not longer exist has, to the extent that it depends on us, a negligible affective significance. Concerning the future, note that the belief that there cannot be happiness except in the *present anticipations* of future events can be reinforced by the conviction that there is literally nothing after the present experience. Consequently, focusing on the present emotions, whatever they are, is more rational than expecting or fearing something yet to come that as of now is nothing at all.

4. Ancient, Modern and Contemporary Examples of Affective Eternalism

In this section, and very schematically, I will distinguish three related ways of characterizing the affective consequences of ontological eternalism, the first focusing on the immense temporal extension of the cosmos as defended by the Stoic philosophy, the second on the difference between imagination and reason in Spinoza's epistemology, and the third on a novel way to cash out the practical and even "ethical" significance of eternalism due to Bertrand Russell. The choice of these three case studies (Stoicism, Spinoza and Russell) is to a significant extent arbitrary, but my extremely brief treatment should be conceived as paradigmatic of related ways in which affective eternalism has been proposed.

4.1 Stoicism on the Sheer Immensity of Time

A characteristic trait of Stoicism is given by the fact that *physics* plays a decisive role both in what we could call—with a little pinch of anachronism—eternalism and in the consequent attainment of wisdom. The wisdom is question consists in the affective attitude that helps us *accept* whatever event the fate (the laws of nature) has prepared for us in the present, given that the event is a consequence of an immensely long (or even eternal) deterministic chain of events, none of which can be avoided. In order to achieve wisdom, the Stoics invite us to contemplate the "rational and necessary unfolding of cosmic events", as Hadot puts it (1995: 59), which is the expression of the lawlike order of the cosmos. The practical eternalist's creed as expressed by ancient Stoicism stresses the fact that it is only by looking at things from a cosmic perspective, and therefore by becoming aware of our insignificant spatiotemporal size with respect to the im-

mensity of time (and space we could add) that can we assign the events that we experience in the present their correct place in the cosmic tapestry.

The reader will excuse this long quotation from Marcus Aurelius' *Meditations*, which I report in full before my critical evaluation because, to my knowledge, it is one of clearest expressions of an ancient philosopher's appeal to look at our life under the aspect of eternity (Ismael's "temporal transcending" view of our life, or Spinoza's expression *sub specie aeternitatis*):

> Think of the whole of being, in which you participate to only a tiny degree; think of the whole of eternity, of which a brief, tiny portion has been assigned to you; think about fate, of which you are such an insignificant part. […] You have the power to strip off many superfluous things that are obstacles to you, and that depends entirely upon your value-judgements; you will open up for yourself a vast space by embracing the whole universe in your thoughts, by considering unending eternity, and by reflecting on the rapid changes of each particular thing; think of how short is the span between birth and dissolution, and how vast the chasm of time before your birth, and how the span after your dissolution will likewise be infinite (Quoted in Hadot 1995: 183).

The affective acceptance of whatever happens to us (loss of health, riches, reversal of fortune, our diseases and upcoming death etc.) is a consequence of our tragic coming to know that "the way things are now" depends on a previous state of the universe, a kind of knowledge that frightens us exactly in virtue of its eternalistic and deterministic[18] ontological basis. By provoking a (slow) psychological change in our immediate emotional reactions, the Stoic form of practical eternalism aims at engendering a different, more adaptive affective attitude toward all the events and objects of our life. This attitude helps us to achieve a rational evaluation of their real importance, which, in turn, is a consequence of our capacity to understand their unavoidable causes. Our coming to know these causes entails the typical anti-anthropomorphic attitude of Stoic philosophy: events *in themselves* are neither good nor bad, they are good and bad only *in relation to ourselves*, that is, relatively to our subjective evaluations. The Stoic eternalism as exemplified by Marcus Aurelius' passage above implies that it is within the limits of our nature to try to control these evaluations themselves, and thereby minimize the dysfunctional emotions that are generated by our interpretation of our experience of external and internal objects. As we are about to see, Spinoza similarly argued that we can replace such anthropomorphic emotions with the joy of understanding the laws of nature, holding everywhere and everywhen, in what today we could call an immutable block universe.

4.2 Spinoza's Epistemology and God's Eternal Laws of Nature

A crucial development of the Stoic affective stance is to be found in Spinoza's *Ethics*, which derives its practical eternalism from an ontology based on eternal, *deterministic laws of nature* leaving no room for teleology or purposes. The eternalism of Spinoza's view of time is justified by his claim that the distinction between past present and future is a by-product of our most imperfect, *first* form of

[18] Here I identify somewhat anachronistically determinism with fatalism, but I think that the identification does not change the meaning of the quotation.

knowledge, that is, *Imagination*. Imagination is filled with purely contingent ideas and cannot therefore apprehend the necessary, eternal laws of nature that Spinoza identifies with his impersonal God and that can be grasped only by *Reason*. Reason is the faculty that is capable of understanding substances in terms of their true, necessary causes: "It is in the nature of reason to perceive things under a certain form of eternity (*sub quâdam æternitatis specie*)".[19] Imagination is a source of error since the imagined ideas are not adequate representations of the essential properties of bodies, which are captured only by our knowledge of laws of nature.

More precisely, this *second* type of knowledge is produced by ideas that are adequate to reality, since they grasp the absolutely necessary, nomological, eternal order of the universe, that is, the immutability and eternity of the laws of nature. For this reason, it is not so anachronistic to attribute to Spinoza an ontological view in which present events are regarded as being on a par with all events constituting the past and the future development of the universe, the latter being governed by eternal nomological relations. Spinoza identifies such laws with an immanent God, *Deus sive natura* (God or nature). Once our reason understands that our mental and corporeal attributes are ruled by the nomological structure of the universe, we achieve the highest, third form of knowledge, which is at the same time an affective attitude toward God/nature, which Spinoza dubs *amor intellectualis Dei* (intellectual love of God). *Prima facie* this expression reads like an oxymoron, involving as it does the *emotion* of love and the intervention of the *intellect*, which is the faculty of reason that is capable to discover the laws of nature.

This impression of conflict is apparent. First of all, Spinoza's impersonal God (*Deus*) is identical with the whole web of the deterministic natural laws, which is accessible only through the intellect (or reason). The emotion of love for the eternal web of laws and therefore for God as he interprets it is the most important affective consequence of Spinoza's eternalism. Spinoza holds that *joy* is a passage from a mental state in which we have less power of acting (we are more passive and less capable of self-preservation) to one in which we have a greater power of acting, and love is simply our *becoming aware* of this passage. Since we naturally love everything that causes this transition, the discovery that the most perfect transition is our coming to know that the necessary laws coincide with the essence of God brings about our intellectual love of nature/God. *The awareness that we can understand Nature brings about love because it is the purest form of activity of human beings.* In sum, the consequence of our coming to know the eternal laws does not bring about a passive attitude of resignation to the destiny, but a joyous, active awareness to belong to something (God) that is either coinciding with the whole cosmos, or is outside time altogether.

4.3 Russell and the Ethical Counterpart of Ontological Eternalism

A much later example of a philosopher holding that the neutrally and impartial outlook of ontological eternalism (my third example) is the key to virtues like altruism and selflessness can be found in Russell's *Mysticism and Logic*:

[19] "It is in the nature of reason to perceive things truly (II. xli.), namely (I. Ax. vi.), as they are in themselves—that is (I. xxix.)". See Spinoza 1996: 59-60.

> The felt difference of quality between past and future, therefore, is not an intrinsic difference, but only a difference in relation to us: to impartial contemplation, it ceases to exist. And impartiality of contemplation is, in the intellectual sphere, that very same virtue of disinterestedness which, in the sphere of action, appears as justice and unselfishness. Whoever wishes to see the world truly, to rise in thought above the tyranny of practical desires, must learn to overcome the difference of attitude towards past and future, and *to survey the whole stream of time in one comprehensive vision* (Russell 1917: 22, my emphasis).

Let us make explicit the link that Russell establishes between an eternalist ontology and an affective attitude of justice and unselfishness. By implicitly endorsing a tenseless ontology, Russell maintains that our intellectually motivated belief that the instant that we *now* occupy in the vast temporal extent of nature is merely perspectival, relational and spatiotemporally located generates the eternalist affective attitude, that Russell refers to as an impartial, allocentric contemplation of the whole stream of nature. In its stress of the indexical nature of "now" and "I", Russell's analogy is very important: the concept of "now" and "self" are strictly related, since the *self* is always situated in a particular moment of *time* (and in a particular location in space) and cannot but look at the world from that a egocentric temporal perspective (Ismael 2007). This perspective is psychologically correlated to the attitude of discounting the future and forgetting the lessons of the past, which can imply to a certain extent our being careless about the continuity of our future selves and that of the others and the future generations, which makes us lose the sense of justice to which Russell refers.

For reasons of space, here I cannot provide more evidence for the historical importance of the connection between the ontological and the practical aspects of presentism and eternalism. In order to strengthen the case in favor of the claim that an independently acquired belief in one of the two ontologies can make a practical difference and see whether it holds water, it is indispensable to clarify the meaning of "live" and "look" in the expressions "live in the present" and look at the world "sub specie aeternitatis".

5. What Does it Mean to "Live in the Present"?

I have already raised an important objection to one of the main claims of the paper, which here is appropriate to formulate in a different way. Since all of our experiences occur in the present, it could be objected that both presentists and eternalists try to avoid as much as possible *present* pain and improve *present* happiness. Therefore, also the eternalists focus on the present experience. However, if I am right, the crucial point is that they try to achieve their common objective in different ways. The different psychological attitude toward the present moment was already clearly formulated in the Hellenistic times. As Hadot notices, the Epicureans' presentism led them to experience the present moment by a *distension* of the mind that, independently of its joyous components, attends to all its contents. On the contrary, the Stoics (who defended eternalism), pursued their aim by a constant *tension* of the mind toward an absolute or partial *control* of the momentary passions. As already noticed, such a tension was possible thanks to the affectively tinged acceptance of the present moment as a necessary consequence of an eternal, lawful order.

Despite the fact that in *daily life* it is impossible both for the presentist not to plan or to think about past experiences in order to avoid future pains and for the eternalist not to attend to her present experiences, the previous quotations have made abundantly clear that the Epicureans more or less explicitly believed that a dynamical form of presentism brings with itself the rationality of a practical attitude that strives for an increased capacity to be absorbed by, care for, and concentrate in, the events that are happening around us. Reinterpreting their claims in contemporary terms, the practical rule "live in the present" means that the mental acts of the presentists are intentionally directed toward the present, in such a way that the experienced events are appreciated for their own sake, as in aesthetic contemplation, scientific creation, deep conversation and play, which are the paradigmatic activities in which present memories of past events and present expectations of future events either play a minor role or no role at all.

More in general, the affections characterising practical presentism are based on the fact that the more we regard the activities in which we are currently engaged as end in themselves, the more meaningful and rewarding they are with respect to the activities that are merely instrumental to reach some other end. Any present activity that is pursued in the present as a mere instrument to reach some future goal is *future-oriented* and the corresponding mental events are intended toward the future. On the contrary when in the present we are engrossed in doing something for its own sake, nothing else in the past or in the future matters.

In the *Nichomachean Ethics* Aristotle put forward a very effective argument in favour of this view, based on the fact that ends are superior to the means that we use to reach them. We can decide to change a means to reach our end, while leaving the latter unchanged: "an end, pursued by itself […] is more complete than an end pursued because of something else […] and an end that is always [choice worthy] and choice worthy in itself, never because of something else, is unconditionally complete" (*Nich*. *Eth*., Book i, 1097a30, transl. in Aristotle 1985: 14).[20] Doing something instrumental to an end presupposes that the end is an effect of the present action.

It might even be suggested that being absorbed in activities for their own sake takes us "outside of time" (the experience of timelessness referred to Wittgenstein in the previous quotation) but what is meant by this provocative expression is that in activities of the kind mentioned above, our awareness of past and future events is somewhat suspended. In a word, when we are mentally engrossed in an activity for the sake of it, we paradigmatically live as affective presentists, that is, as accomplishers and realizers of an ontological doctrine. It seems safe to conclude that a dynamical form of ontological presentism can *de facto* and ought to be an important motivator for the effort of making the most of what we are experiencing "right now".

The description in some more detail of the affective consequences of adopting the motto "look at your life sub specie aeternitatis" will also give me the opportunity to raise two additional objections to the main claim of the paper.

[20] See also Schlick 1987 and Russell: "there can be no value in the whole unless there is value in the parts. Life is not to be conceived on the analogy of a melodrama in which the hero and heroine go through incredible misfortunes for which they are compensated by a happy ending" (Russell 1930: 24).

6. What Does it Mean to "Look at One's Life Sub Specie Aeternitatis"?

Going back to the eternalist's affective outlook described by Marcus Aurelius' quotation above, it is clear that he was supposing that a belief in ontological eternalism could generate the affective belief that, in order to answer in an emotionally appropriate way the challenges of our present experiences, we must temporally locate them in the complete history of the universe. A more credible and less radical formulation of affective eternalism could just consists in the claim that our frequent reflection on the true physical description of our spatio-temporal place in the cosmos could help us to avoid a dramatization of "*relatively small*" setbacks or complications of the present moment (say, missing a plane or arriving second to a race, or suffering a theft, etc.), by realizing in addition that, *qua* consequences of a long chain of events preceding our life, they have a negligible meaning.

The first objection amounts to a dilemma: a less radical but more reasonable formulation of affective eternalism is uninteresting, the more coherent one is impossible to achieve. The expression "relatively small" of the previous paragraph is ambiguous: suffering a theft can be "nothing" also for an affluent presentist but can be tragic for a poor eternalist. However, the coherent eternalist should react to all events (even the most tragic ones) in the same way, by locating them in the temporal vastness of the cosmos. It seems clear that this is an impossible ideal because the death of one's son cannot be compared to losing one's wallet.

The thesis implied by the first horn of the dilemma seems highly controversial: thinking that the universe is 13,4 billiards year old can be of help in many practical circumstances, even we cannot describe them one by one. As far as the second horn is concerned, the eternalist can reply by pointing out that affective eternalism can *help* to accept also the most tragic events by reflecting on her metaphysical assumptions. She needs not be fatalistic: dwelling on a terrible *present* tragedy is inevitable but projecting one's life in an existing future is the only way to make the present more bearable. This option is open only to eternalists. In 1980, after an earthquake in the southern part of Italy that caused many casualties the past president of the Italian Republic Sandro Pertini said: "the best way to remember those who are dead is to think about those who are still alive". And a large part of those who are alive, for an eternalist but not for a presentist, have and will have a future, and we can make a difference to make it better, People have reported than thinking about Pertini's words after the tragedy helped them to suffer less.

The second objection is that the eternalist attitude would make *any* moment of our life *utterly insignificant.* The response here is that a well-grounded belief in ontological eternalism could create a sense of solidarity and compassion for our fellow beings and all living beings sharing the tragic destiny of death and pain with us all in a temporal immense universe. This emotional attitude, defended in particular by Schopenhauer (1958), can be endorsed without subscribing to his metaphysical irrationalism, based on the belief of a blind *Wille* (Will) hidden behind the veil of our *Vorstellung* (representation of phenomena). The creation of a strong tie of solidarity and compassion among human beings is a plausible consequence of the awareness of the brevity and impotence of our life if compared with the immense temporal size of the universe.

It should be noted that, beyond voluntary reflections on eternalism, the corresponding attitude is typically and implicitly stronger in scholars dedicating their carrier to, the study of cosmology, astrophysics, geology, evolutionary biology and, to a minor extent, human history. All of these disciplines can be instrumental to adopt a more detached and allocentric attitude toward our present experience.[21]

By zooming in from the temporal length of human history to the length of our own life, ontological eternalism also implies the belief that all events of our life are ontologically on a par. The corresponding affective attitude toward our existence then becomes correlated to an important question that here cannot be discussed but that must at least be mentioned, namely the "constitution" of the self as an entity that is extended in time (Korsgaard 2009; Ismael 2016). The eternalist's typical emotional stance motivates the belief that each action and decision taken in the present moment must be part of a coherent narrative that ought to guide our selves during our entire life. The future is going to be affected by the present decisions, which must also *cohere* with those actions and values that have inspired our life. This coherence need not include only events between our birth and death but pushes us to extend our ethical interests also to events preceding and following our life. As far as the past is concerned, for instance, the affective stance following from ontological eternalism may help us to extend our care also to the legacy of previous generations, especially when it is characterized by the attempt to achieve social and cultural ends, like the advancement of knowledge and the extending to all mankind the right of living a dignified life and receiving an education. Likewise, the eternalist affective attitude can stimulate the obligation of focusing our actions also to the future generations. In this more extended sense, the coherent narrative that a practical existentialist tries to achieve in her own *individual* life must be extended to the past and future generations as well, in order to bequeath the best ideal of the former to the latter.

In the previous part of the paper, I did not clarify the relationship between a psychological affective consequence of ontological presentism or eternalism and a pragmatic rule that should guide our concrete actions and could follow from the beliefs in the two ontologies. How can a rational constraint on our actions follow from our belief in the two ontological views with their respective affective consequences?

7. An Objection to the Practical Importance of the Two Ontological Views

In order to tackle this issue, I must discuss a key objection to my main thesis. Recall the argument presented above:

(i) The ontological dispute is non-verbal.
(ii) There are two different affective attitudes related, respectively, to ontic presentism and eternalism that are influenced only by our purely epis-

[21] Such attitudes are made possible by a mechanism called mental time travel, which recently has been object of intense neurocognitive studies, and which consists in the capacity to stretch one's imagination to more or less long temporal intervals. See among others Suddendorf et al. 2009; Arzy et al. 2016. Buonomano 2017 is an elementary ex-position. It turns out that the capacity to create allocentric spatial maps is mirrored by that of creating allocentric temporal maps, by which we get in "cognitive contact" with future (and past) events from the perspective of our present experience. For a nice, brief review of the difference between egocentric and allocentric temporal maps, see Callender 2017: 207-20.

temically motivated beliefs in the two respective ontologies: *namely affective presentism and affective eternalism*.

(iii) Our *actions* are at least partly motivated by our affections or emotions.

Conclusion: these different attitudes toward time make some *practical difference* in how we act (and possibly should act).

The objection points out that an epistemic commitment to one of the two ontologies need not have a *univocal* affective consequence. The objection can be stated thus:

(iv) A believer in *ontological presentism* can look at the world *sub specie aternitatis* as well and often as the *practical eternalist*. Conversely, a believer in *ontic eternalism* can as well and as often be completely engrossed in her present experience like a *practical presentist*.

In a word, (iv) does not deny that an epistemic commitment to one of the ontologies can have affective consequences. It just affirms that a (ontic) presentist can *in different moments of her* life be a pragmatic eternalism and conversely, without abandoning her epistemic commitment to the respective ontic view. It follows that the different ontological commitments to presentism (eternalism) do not suffice to fix the respective affective stances, since the same affective stance is compatible with the two different ontological commitments. If the two ontologies are *underdetermined* by the affective stances, it seems plausible to conclude that a belief in one of the two ontologies does not make a *temporally stable difference* in her affective attitudes, so that it does not make any important pragmatic difference. "Temporally stable" here refers to a prevalent *character trait* that is reinforced by an epistemic belief in one of the two ontologies: the italicized expression will be clarified and become important in what follows.

Objection (iv), if correct, would have two important consequences.

The first, if (iv) is correct, is relevant to Deng's (2017) hypothesis that the presentist/eternalist dispute is merely verbal and "reduces simply" to an affective dimension. If, as Deng has it, (i) is false, we could not conclude with her that the eternalist/presentist debate reduces *simply* to the two different and incompatible affective attitudes towards time. As a consequence of the underdetermination thesis, and even dropping as she does any reference to ontological claims, the "reduction" in question is much more complicated than could be expected.

On the other hand, by accepting premise (i), as I have done here, the importance and the interest of the affective dimensions of the debate illustrated above would be even greater, even if (iv) were correct. It is only if we accept the genuine character of the debate that the historical positions that we briefly commented above could be explained.

The *second* consequence amounts to a rejection of (iv) on the basis of the pragmatist view that our beliefs are guides to actions: whatever makes some practical difference ought to make some epistemic difference. However, the first consequence claims that our beliefs in the two ontologies make no pragmatic difference because they make no *temporally stable* affective difference. If a pragmatist's initial trust in the fact that believing in one of the ontologies could have practical consequences were followed by the discovery that there is no epistemic difference between the two ontologies, she might plausibly end up in a state of *epistemic neutrality*. Such a neutrality between the two ontologies, which is com-

patible with the truth of (i), would bring with itself *indifference*, an additional affective consequence not contemplated before, but that is typical of an anti-metaphysical philosophical position.

The real way out of (iv) relies on James's notion of *temperament* (James 1979: 7). By invoking this notion, the independence of one's affective attitudes from one's belief in one of the two ontologies claimed by (iv) would be substantially weakened. *In fact, it would be undermined by the claim that an epistemic commitment to a given ontology influences or reinforces a previously present, temporally stable temperament or character*. This claim is all that is needed to defend the conclusion of the three premises above and therefore the claim that a rational, epistemic motived commitment to one of the two ontologies reinforces the affective stance of a distinct kind of "time oriented" person.

Just to illustrate, James refers to "the realist philosopher" as a tough-minded person and the idealist philosopher as a tender-minded person (see 1979, *ibid*.).[22] By modifying his distinction in order to apply it to our case, the physicalist outlook that attracts the tough-minded philosopher and that inspires her eternalism thrives on the idea that one of the aims of metaphysics, science and physics (recall the Stoic position) is a sort of liberation from our anthropomorphic beliefs, of which ontological presentism is a fundamental ingredient. On the opposite side, the "tender-minded" ontological presentist wants a universe in which not only is our experience of an objectively privileged present veridical, but it even takes precedence over the physicalist, eternalistic outlook, independently of any evidence physics may have in its favor. Consequently, despite our negligible place in the large scheme of things, the temperament of ontological presentists (eternalist) pushes towards the adoption of ontological presentism.

This pragmatist outlook, however, should not be generalized to the point of endorsing James' general claim that "the history of philosophy is to a great extent that of a certain clash of human temperaments" (James 1979: 7). In philosophy, the *initial motivation* to adopt a certain metaphysical position may depend on our character trait, but must be justified *only* by rational arguments that can be brought in its favor, and therefore, in our case, not by Jamesian "time-related" temperaments or affective stances. The point that I am stressing here is that, given the presence of stable character traits, (iv) can be weakened if not rejected by the claim that temporally stable character traits will be reinforced by epistemically motivated commitments to one of the ontologies. This commitment can at the same time independently reinforce the values she cherishes most in virtue of her character trait, even though these values do not play any role in logically justifying her position.

8. Conclusions

Despite the reasonable defense of eternalism given above, it must be admitted that a commitment to this ontology can change our emotional reactions only to a certain extent, even if it can make an important difference. This was explicitly

[22] To cut James's description short, tough-minded philosophers stick to fact, i.e., are realist, pessimistic, and irreligious. The tender minded are idealistic, optimistic, religious, and free-willist (1979, *ibid*.).

recognized by a well-known defender of eternalism: in a much less famous letter of condolence sent to the mathematician Elie Cartan on May 21 1930,[23] Einstein openly claims that an intense suffering in the present is to some degree *irrational* because objectively there is no now: "In these trying moments one feels how it is difficult for a human being to hold fast to the idea—so inescapable to a physicist—that the now is only an illusion, not something pertaining to reality". In this passage Einstein seems to be implying that in less trying moments, a firm belief that the present has no objective existence should make our pain more tolerable since our temporal experience amounts just to arbitrary perspective on the immense temporal and spatial extension of the universe. Cultivating eternalist thoughts and, consequently, affective attitudes of this kind is a different way a vindicating a fact already insisted upon by the Stoics: the adoption of a particular ontological view of time makes and can make an important difference in how we should live.[24]

References

Annas, J. 1995, *The Morality of Happiness*, Oxford: Oxford University Press.

Aristotle 1985, *Nichomachean Ethics*, transl. T. Irwin, Cambridge, MA: Hackett.

Arzy, S. et al. 2009, "Subjective Mental Time: The Functional Architecture of Projecting the Self to Past and Future", *European Journal of Neuroscience*, 10, 2009-17.

Brink, D. 2011, "Prospects for Temporal Neutrality", in Callender, C. (ed.), *Oxford Handbook of Philosophy of Time*, Oxford: Oxford University Press, 353-81.

Bradley, B. 2004, "When Is Death Bad for the One Who Dies?", *Noûs*, 38, 1-28.

Buonomano, D. 2017, *Your Brain Is a Time Machine: The Neuroscience and Physics of Time*, New York: W.W. Norton.

Burley, M. 2008, "Should A B-Theoretic Atheist Fear Death?", *Ratio* 21, 3, 260-72.

Callender, C. 2017, *What Makes Time Special*, Oxford: Oxford University Press.

Casati, R. and Torrengo, G. 2011, "The not-so Incredible Shrinking Future", *Analysis*, 71, 2, 240-44.

Cockburn, D. 1997, *Other Times. Philosophical Perspectives on Past Present and Future*, Oxford: Oxford University Press.

Correia, F. and Rosenkranz, S. 2018, *Nothing To Come: A Defence of the Growing Block Theory of Time*, Cham: Springer.

Crisp, T. 2004, "On Presentism and Triviality", in Zimmermann, D. (ed), *Oxford Studies in Metaphysics*, Vol. 1, Oxford: Oxford University Press, 15-20.

[23] The more famous letter I am referring to was sent to Besso's sister after Besso's death, in which he famously writes that for believing physicists the difference between past present and future is only a stubborn illusion. In my opinion, the letter to Cartan states in a much clearer way why Einstein thought that a firm belief in eternalism "*rationally*" should, even if it actually may not, alleviate a present pain.

[24] I thank the anonymous referees for their comments and the editor-in-chief for his precious help. This work was supported by the Italian Ministry of Education, University and Research through the PRIN 2017 program "The Manifest Image and the Scientific Image" prot. 2017ZNWW7F_004.

Deng N. 2017, "What is Temporal Ontology?", *Philosophical Studies*, 175, 793-807; DOI https://doi.org/10.1007/s11098-017-0893-6

Dolev, Y., 2007, *Time and Realism: Metaphysical and Antimetaphysical Perspectives*, Cambridge, MA: The MIT Press.

Dorato, M. 2008, "Putnam on Time and Special Relativity", *European Journal of Analytic Philosophy*, 4, 2, 51-70, with a "Reply to Mauro Dorato", by Hilary Putnam, 70-73.

Green, P. and Sullivan, M. 2015, "Against Time Bias." *Ethics*, 125, 1-24.

Finocchiaro, P. and Sullivan, M. 2016, "Yet another 'Epicurean' Argument", *Philosophical Perspectives*, 30, 135-59.

Hadot, P. 1995, *Philosophy as a Way of Life: Spiritual Exercises from Socrates to Foucault* (ed. with an introduction by A. Davidson, transl. from French by M. Chase), Oxford: Blackwell.

Harrison, J. 1981, *Horace in his Odes* (Latin text, ed. and transl. from Latin by J. Harrison), Wauconda, IL: Bolchazy-Carducci.

Horace 2004, *The Works of Horace* (trans. literally into English Prose by C. Smart, a new edition revised by T. Buckley B.A. of Christ Church): https://www. gutenberg.org/files/14020/14020-h/14020-h.htm

Inwood, B. and Gerson, L. (eds.) (1994), "Letter to Menoeceus", in *The Epicurus Reader*, Indianapolis, IN: Hackett.

Inwood, B. and Gerson, L. (eds.) (1997), *Hellenistic Philosophy*: *Introductory Readings* (2nd edition), Indianapolis, IN: Hackett.

Ismael, J. 2007, *The Situated Self*, Oxford: Oxford University Press.

Ismael, J. 2016, *How Physics Makes Us Free*, Oxford: Oxford University Press.

James, W. 1979, *Pragmatism*, Cambridge, MA: Harvard University Press.

Korsgaard, C. 2009, *Self-Constitution*: *Agency, Identity, and Integrity*, Oxford: Oxford University Press.

Le Poidevin, R. 1995, "Time, Death and the Atheist", *Cogito*, 9, 2, 145-52.

Maudlin, T. 2007, "On the Passing of Time", in *The Metaphysics within Physics*, Oxford: Oxford University Press, 104-42.

Merricks, T. 2007, *Truth and Ontology*, Oxford: Oxford University Press.

Mellor, D. 1999, "Review of 'Other Times: Philosophical Perspectives on Past, Present and Future' by David Cockburn", *The Philosophical Review*, 108, 3, 428-30.

Meyer, U. 2005, "The Presentist's Dilemma", *Philosophical Studies*, 122, 213-25.

Mozersky, J. 2011, "Presentism", in Callender, C. (ed.), *The Oxford Handbook of Philosophy of Time*, Oxford, Oxford University Press, 122-44.

Norton, J. 2014, "The Burning Fuse Model of Unbecoming in Time" *Studies in History and Philosophy of Modern Physics* 52, 103-05.

Orilia, F. 2016, "On the Existential Side of the Eternalism-Presentism Dispute", *Manuscrito Rev. Int. Fil. Campinas*, 39, 4, 225-54.

Pooley, O. 2013, "Relativity, The Open Future and the Passage of Time", *Proceedings of the Aristotelian Society*, CXIII, 3, 321-63.

Rosenbaum, S. 1989, "The Symmetry Argument: Lucretius Against the Fear of Death", *Philosophy and Phenomenological Research*, 50, 2, 353-73.

Russell, B. 1917, *Mysticism and Logic*, London: George Allen & Unwin.

Russell, B. 1930, *The Conquest of Happiness*, New York: Norton and Co.

Savitt, S. 2006, "Presentism and Eternalism in Perspective," in Dieks, D. (ed.), *The Ontology of Spacetime*, Vol. 1, Amsterdam: Elsevier, 111-27.

Schlick, M. 1987, "On the Meaning of Life", in Mulder, H. and van de Velde-Schlick, B.F. (eds.), *Philosophical Papers*, Dordrecht: Reidel.

Schopenhauer, A. 1969, *The World as Will and Representation*, Vols. I and II (transl. by E.F.J. Payne), New York: Dover Publications.

Setiya, K. 2014, "The Midlife Crisis", *Philosopher's Imprint*, 14, 31, 1-18.

Sider, T. 2016, *Four-Dimensionalism*, Oxford, Clarendon Press.

Silverstein, H. 1980, "The Evil of Death", *The Journal of Philosophy*, 77, 7, 401-24.

Skow, B. 2015, *Objective Becoming*, Oxford: Oxford University Press.

Spinoza, B. 1996, *The Ethics* (ed. and transl. by E. Curley), London: Penguin Books.

Suddendorf, T. et. al. 2009, "Mental Time Travel and the Shaping of the Human Mind", *Philosophical Transactions of the Royal Society B* 364, 1317-324, DOI: 10.1098/rstb.2008.0301.

Sullivan, M. 2018, *Time Biases*, Oxford: Oxford University Press.

Weatherall, J. 2020, "Essay Review of 'What Makes Time Special?' by Craig Callender", *Philosophy of Science*, 87, 3: DOI: https://doi.org/10.1086/709118

Wittgenstein, L. 1961, *Tractatus Logico-Philosophicus*, transl. by D.F. Pears and B.F. McGuinness, New York: Humanities Press.

A Note on the Grandfather Paradox

Brian Garrett

The Australian National University

Abstract

In this note, I am critical of some aspects of David Lewis's resolution of the Grandfather Paradox. In particular, I argue that Lewis gives the wrong explanation of Tim's inability to kill Grandfather, and that the correct explanation makes essential reference to the self-undermining character of Tim's grampicide.

Keywords: David Lewis, Time travel, Grandfather paradox.

The philosophy of time travel is, in large part, an attempt to answer the exam question: To what extent, if any, do you disagree with the views defended by David Lewis in his eminently readable "The Paradoxes of Time Travel"? (Lewis 1976). One of the most interesting and influential parts of Lewis's article is his discussion of what a traveler to the past can and can't do. In particular, can such a traveler kill his own grandfather?

In Lewis's thought-experiment, we are asked to consider Tim, who evidently dislikes his grandfather, and has built a time machine in order to go back to 1920 and kill him, many years before Tim's father was conceived. Tim duly travels back to 1920, buys a rifle, and tracks the route of Grandfather's daily walk (Lewis 1976: 149).

According to Lewis, Tim can kill Grandfather. He has a high-powered rifle; he's a good shot; weather conditions are perfect, etc. On the other hand, Tim can't kill Grandfather. Grandfather died in his bed in 1957. Consistency demands, despite his best efforts, that Tim fail in his attempt to kill Grandfather, and fail for some commonplace reason (an errant seagull, a distracting noise, an observant policeman, etc.) (Lewis 1976: 150).

Since Lewis holds that Tim can and can't kill Grandfather, it might be thought that his position is contradictory. Lewis has a nice reply to this charge. There is no contradiction since 'can' is context-dependent. He writes:

> To say that something can happen means that its happening is compossible with certain facts. Which facts? That is determined, but sometimes not determined well enough, by context. An ape can't speak a human language, say, Finnish, but I can. Facts about the anatomy and operation of the ape's larynx and nervous system are

Argumenta 9, 1 (2023): 191-193
ISSN 2465-2334

© 2021 Brian Garrett
DOI 10.14275/2465-2334/20210.gar

> not compossible with his speaking Finnish. The corresponding facts about my larynx and nervous system are compossible with my speaking Finnish. But don't take me along to Helsinki as your interpreter: I can't speak Finnish. My speaking Finnish is compossible with the facts considered so far, but not with further facts about my lack of training. What I can do, relative to one set of facts, I can't do, relative to another, more inclusive, set (Lewis 1976: 150).

According to Lewis, then, 'can'-judgements are context-dependent. Relative to one context 'A can do F' is true, relative to another it's not. In Lewis's example, given the facts about the structure of my larynx and nervous system, I can speak Finnish. But given a wider set of facts, including the fact that I have never learnt Finnish, I can't. Similarly, given one set of facts, e.g., facts about Tim's rifle, his shooting ability, the weather conditions, etc., Tim can kill Grandfather. But given another, more inclusive, set of facts, including, e.g., the fact that Grandfather wasn't killed in 1920, Tim can't kill Grandfather.

Let's concede to Lewis that there's a sense in which Tim can kill Grandfather. Relative to facts about Tim's means, motive and opportunity, Tim can kill Grandfather. Here I want to focus on the sense in which Tim can't kill Grandfather, and on what Lewis has to say about it. Lewis writes: "Tim cannot kill Grandfather. Grandfather lived, so to kill him would be to change the past" (Lewis 1976: 150). It is, as Lewis rightly notes, logically impossible to change the past (or the present or the future). No one can make it the case that some event which didn't happen did or that some event which did happen didn't.

Unfortunately, the second sentence in the quote from Lewis does not support its first sentence. The impossibility of changing the past implies only that, since Grandfather wasn't killed in 1920, Tim won't kill him then. It doesn't imply that Tim can't kill him then.[1] Indeed, Lewis seems to be endorsing the invalid inference pattern: $\sim \Diamond(A \,\&\, \sim A)$; $\sim A$; so $\sim \Diamond A$. That is: Tim can't both kill and not kill Grandfather; Tim doesn't kill Grandfather; so Tim can't kill Grandfather.

However, there is a sense in which Tim can't kill Grandfather, but its ground is not the impossibility of changing the past. Its ground is rather the fact that Tim's homicide is self-undermining. As Lewis observes: "No Grandfather, no Father; no Father, no Tim; no Tim, no killing" (Lewis 1976: 152). A self-undermining action is one which undermines a causally necessary condition for its agent's existence in the first place. (Suicide, of course, is not a self-undermining act in this sense.) Plainly, no agent can perform a self-undermining act. That is, no agent can make it the case that he never existed. Indeed, a self-undermining act, so defined, would seem to be logically impossible since its performance requires both that its agent exist and never existed.

In terms of Lewis's context-dependent theory, we can put the point as follows: relative to the fact that Tim's action is self-undermining, Tim can't kill Grandfather. Kadri Vihvelin also holds Tim can't kill Grandfather, on the grounds that no matter how often Tim tried to kill Grandfather, he would fail.

[1] Romy Jaster, in her contribution, also fails to vindicate any sense in which Tim can't kill Grandfather. According to her version of the context-dependent view of 'can'-judgements, Tim can't kill Grandfather because "[he] does not shoot [Grandfather] in a sufficient proportion of the possible situations in which he intends to shoot him *and the fact that he does not shoot him obtains*" (Jaster 2020: 104; italics in text). Since this sentence is trivially true it can hardly imply that Tim can't kill Grandfather (trivialities only imply trivialities).

(Vihvelin 1996; 2020) The account offered here is preferable since it explains why Vihvelin's conditional is true (*viz.*, no one can perform a self-undermining action).

Given the preceding discussion, we can see that Lewis is wrong to urge a complete parallel between Tim and Tom. Tom is a normal (non-time travelling) inhabitant of 1920. He wants to kill Grandfather's partner, who lives until 1960. Tom will of course fail in his attempt since we have stipulated that Partner die in 1960. Thus, Tim and Tom are alike to the extent that each will fail in their homicidal attempts. However, Lewis says that Tom can't kill Partner for the very same reason that Tim can't kill Grandfather: *viz.*, neither man was killed in that year, and no one can change the past, present or future (Lewis 1967: 151). As we have seen, this reasoning is flawed. Furthermore, since Tom's action, unlike Tim's, is not self-undermining, nothing stands in the way of Tom killing Partner in 1920 (given that he has the means, motive and opportunity). The whole truth about Tom's situation is: he can kill Partner, but he won't.

The same is true of non-self-undermining attempts to undo the past. Presumably, many actions that Tim can (but doesn't) perform in 1920 wouldn't undermine his own existence. In these cases, descriptions of the 'can but won't' (and not the 'can't') variety apply. For example, suppose that Tim never shook Grandfather's hand in 1920. Can Tim shake his hand then? Yes, he can, although he won't. Indeed, to think otherwise—to think that Tim can't shake Grandfather's hand because it wasn't shaken then—is to succumb to the fallacious reasoning identified above.

In sum, then, I am critical of Lewis's resolution of the Grandfather Paradox on three fronts. First, Lewis gives the wrong explanation of the sense in which Tim can't kill Grandfather. Second, Lewis fails to emphasise the right explanation: Tim can't kill Grandfather relative to the fact that his action is self-undermining. Third, since Tom killing Partner is not self-undermining, Lewis is wrong to press for a complete parallel between Tim and Tom. In Tom's case, there is no fact relative to which he can't kill Partner (assuming that he has the requisite means, motive and opportunity).[2]

References

Jaster, R. 2020, "What Tim Can and Cannot Do: A Paradox of Time Travel Revisited", *Kriterion*, 34, 4, 93-110.

Lewis, D. 1976, "The Paradoxes of Time Travel", *American Philosophical Quarterly*, 13, 1, 145-52.

Vihvelin, K. 1996, "What Time Travelers Cannot Do", *Philosophical Studies*, 81, 315-30.

Vihvelin, K. 2020, "Killing Time Again", *The Monist*, 103, 312-27.

[2] I am grateful to Alan Hájek, J.J. Joaquin, Daniel Stoljar, and an *Argumenta* referee, for useful feedback.

Virtue, Character, and Moral Responsibility: Against the Monolithic View

Giulia Luvisotto* and Johannes Roessler**

* University of Helsinki

** University of Warwick

Abstract

A traditional tenet of virtue ethics is that a proper moral assessment of an *action* needs to be informed by a view of the *agent*; in particular, a view of their virtues or vices, as exhibited in their action. This picture has been challenged on the grounds that it is revisionary and ill-motivated. The key claim is that we are ordinarily disposed to judge the moral merits of particular actions independently of any view of the character of the agent, and that there is nothing wrong with that practice. In this paper, we identify and criticize a certain view of the nature of character that (we argue) underpins the challenge. We call this a monolithic conception of character. We sketch an alternative, non-monolithic conception, and suggest that when combined with a non-monolithic conception, the traditional tenet can be seen to be neither revisionary nor ill-motivated.

Keywords: Virtues, Character, Moral responsibility, Reason, Explanation.

> *We are all patchwork, and so shapeless and diverse in composition that each bit, each moment, plays its own game.*
> (Montaigne 2003: 296)

1. Introduction

Virtue terms are used in two ways: they are applied both to people and to their actions. Suppose that you love inviting friends over for dinner and serve them delicious delicacies, promptly share your research insights with your colleagues, and typically think the best of everyone. In brief, you are a generous person, someone who sees the possibility of sharing as a good reason to do so. You display the property of being generous. Yet, generous is also what you do. Your hosting a sumptuous dinner or sharing your insights were generous actions. As Thomas Hurka puts it, "moral thought uses the concepts of virtue and vice at two different levels", a "global" and a "local" one (Hurka 2006: 69).

Argumenta 9, 1 (2023): 195-207
ISSN 2465-2334

© 2022 The Authors
DOI 10.14275/2465-2334/20220.luv

How are the two kinds of uses of virtue terms related to one another? According to a venerable tradition, only actions that (in Aristotle's words) "proceed from a firm and unchangeable [virtuous] character" properly count as virtuous (Aristotle 1980: 1105[a]). A 'local' use of a virtue term in appraising what someone is doing or has done is not, according to this tradition, independent of a 'global' use of the relevant term in thinking about the agent's character. This view is widely seen as partly definitive of a 'virtue ethical' approach to moral philosophy. An assessment of the moral merit of an action is supposed to be *informed* by an assessment of the character traits exhibited by the action (see, for example, Hursthouse 1999, Annas 2007). Whether an action is generous depends on whether the agent is. Call this the Dependence thesis.

Despite (or possibly because of) its venerable pedigree, the dependence thesis can look like a piece of philosophical theorizing that is far removed from the way we ordinarily think about moral responsibility. Critics of the thesis often invoke cases in which an agent, for the first time or anyway 'out of character', performs an action that nevertheless merits the local use of a virtue term. We call the intuition that is supposed to be elicited by this style of reflection the 'single instance intuition'. We are particularly interested in two lessons that have been drawn from that intuition: first, that the Dependence thesis is *revisionary*;[1] second that it lacks a convincing *rationale*. Participants in our ordinary practice of holding each other morally responsible, the claim is, are happy to assess the merits of an action independently of reflection on the 'firm and unchangeable' character traits (if any) from which the action proceeds. And there is no good reason, it is argued, to impugn that practice.

Our aim in what follows is to develop a version of the Dependence thesis that is able to rebut both charges. We grant that the single instance intuition has considerable force, but its interpretation is a delicate matter. We suggest that the intuition only counts against overly rigid versions of the Dependence thesis, versions that assume what we will call a monolithic conception of character. We suggest that the central notion we should appeal to in defending the Dependence thesis is the notion of the agent's 'evaluative orientation', and that this leaves significant latitude regarding the nature of character traits (Section 3). We go on to argue that a (non-revisionary) rationale for the Dependence thesis emerges from reflection on the nature of ordinary reason-giving explanations of actions (Section 4).

2. The Single Instance Intuition

Let us start with some examples intended to elicit the single instance intuition. Here are two cases from Thomas Hurka:

[1] Not all theories of responsibility would deny the Dependence Thesis. Real self views for instance maintain that a person is responsible for an action insofar as it is attributable to their real self, display their values. As Susan Wolf puts it (to introduce the view, which she opposes) "an agent's behavior is attributable to the agent's real self…if she is at liberty (or able) both to govern her behavior on the basis of her will and to govern her will on the basis of her valuational system" (Wolf 1990: 33). However, note that the Real self views are in principle compatible with the monolithic view of character, which is our main point of contention in what follows: they can maintain that an action is attributable to an agent only if it displays *robust and stable* dispositions.

> Imagine that, walking down the street, you see someone kick a dog from an evident desire to hurt the dog just for the pleasure of doing so. Do you say, 'That was a vicious act' or 'That was a vicious act on condition that it issued from a stable disposition to give similar kicks in similar circumstances'? Surely you say the former. Or imagine that your companion stops to give $20 to a homeless person, apparently from concern for that person for her own sake. Do you say, 'That was generous of you' or 'That was generous of you on condition that it issued from a stable disposition to act from similar motives in similar circumstances'? Again surely you say the former (Hurka 2006: 71).

The examples are framed in such a way as to emphasize a contrast between the agent's *current motivation*—something that is supposedly "evident" or "apparent"—and their *stable dispositions* for acting in relevantly similar ways, of which we may be ignorant. The agent, in these examples, may well have a stable disposition to act viciously or generously; the important point is that whether they do seem to be completely irrelevant in the context of a local judgement regarding the moral merit of their action. We can put the lesson we are supposed to draw from such examples like this:

(1) We often take ourselves to be justified in judging an action to be generous, in the absence of any independent evidence that the action manifests a stable disposition for generous behaviour.
(2) We would not ordinarily take a single generous action to provide adequate evidence for crediting the agent with a stable disposition for generous behaviour.
(3) Therefore, we do not ordinarily take the moral merit of an action to depend on the character traits exhibited by the action.[2]

The upshot is that the Dependence thesis is revisionary. Or, as Hurka puts it, more bluntly: "too much attention to ancient philosophy can blind one to what I think are obvious facts about the everyday understanding of virtue" (Hurka 2006: 74).

Consider next an example of Rosalind Hursthouse's (one she discusses as a potential counterexample to the view she is defending): "Someone described as 'absolutely ordinary', 'not courageous at all', suddenly 'uncharacteristically' does something quite heroic" (Hursthouse 1999: 157). In this sort of case, it is not that we are ignorant of the agent's character. We know, or anyway think we know, that they are not courageous, yet we supposedly don't hesitate to contemplate the possibility that they may have acted courageously. The intuition can be pressed further by comparing two examples of a courageous action: one that manifests a stable character trait and one performed 'out of character'. Is there any reason to assume the former is more commendable than the latter? Straight off, it seems this would be akin to saying that the cake that you, a skilled baker, just baked is nicer than the one I just baked, which is the result of my first-ever attempt, just because your cake stems from more developed and reliable skills than mine. But if we followed the same recipe to the letter, used the same ingredients, tools and oven, there is surely no reason to think that my cake is any less delicious. My cake is no less good a cake *qua first attempt*. Seen in this light, the Dependence thesis can

[2] McCormick and Schleifer put the argument succinctly: "Can we really even assess whether someone possesses a particular virtue based on one instance? It seems not, but we can still blame him in this one instance" (McCormick and Schleifer 2006: 79).

seem bewildering. In our ordinary practice of treating each other as responsible agents, what seems to matter is the *motive* informing an action, not the long-lasting character traits the action may exhibit. What would be the rationale for withholding praise from an act, merely on the grounds that it was not 'characteristic'?

Our aim here is not so much to resist the single instance intuition as to probe and unpick the terms in which critics of the Dependence thesis interpret it. Once the intuition has been detached from its misleading interpretation, we suggest, it no longer looks like a challenge to the Dependence thesis: on the contrary, it can play a significant role in developing the thesis.

To see that there are grounds for suspicion about the standard way of framing the single instance intuition, consider Hurka's embellishment of Hursthouse's 'out of character' case. Hurka imagines a military committee entrusted with the decision whether to give a soldier a medal for bravery. He asks:

> Would they say, 'We know he threw himself on a grenade despite knowing it would cost him his life and in order to save the lives of his comrades. But we cannot give him a medal for bravery because we do not know whether his act issued from a stable disposition or was, on the contrary, out of character'? They would say no such thing, and they would be obnoxious if they did (Hurka 2006: 72).

It seems intuitive that 'they would say no such thing', and it seems plausible, moreover, that the point tells us something about 'our everyday understanding of virtue'. Yet note that on Hurka's construal of the distinction between local and global uses of virtue terms, the committee could reasonably be expected to elucidate their decision as follows: 'We know he performed a brave act. That is why we are giving him a medal. We should like to put on record, however, that we are not implying that he is a brave person, or even just a brave soldier. The award reflects our local judgement about the act he performed; it should not be taken to reflect any global assessment of the sergeant himself'. Straight off, this seems no less strange than withholding the medal on the grounds of uncertainty about stable dispositions. And that observation also seems to tell us something about our everyday understanding of virtue. It is not just that it would be churlish to make the distinction between the two kinds of judgement explicit. Rather, we would ordinarily take it to be offkey to separate a local from a global judgement: an award of a medal for bravery is naturally interpreted as amounting to both. Consider the awardee's own reaction: he will be inclined to feel good, surely, not just about what he did but also, connectedly, about who he is.

Are the intuitions generated by Hurka's committee example and by our variation on that example in conflict with each other? We want to suggest that they are not. They can be seen to be mutually compatible by probing and dislodging an unargued assumption that informs Hurka's interpretation of the single instance intuition: the assumption that that the global use of a virtue term amounts to an attribution of a 'firm and unchangeable character trait' or a 'stable disposition'. We call this a monolithic conception of character. The suggestion we wish to explore is that the Dependence thesis is not in fact committed to the monolithic conception. If we discard the latter, we can interpret the single instance intuition in a way that makes it compatible with the Dependence thesis. The basic idea is this: *even if a single instance may not suffice to give us a complete portrait of who someone is and what values they have, it does suffice to tell us something about them as a person.* Thus, while the monolithic conception (we suggest) is indeed revisionary, the

Dependence thesis is not, or at least is not revealed to be so by the single instance intuition. We develop this diagnosis in the next section. In Section 4, we come back to the question of the rationale that might be offered for the Dependence thesis.

3. The Monolithic Conception of Character

Montaigne wrote: "Therefore one courageous deed must not be taken to prove a man valiant; a man who was really valiant would be so always and on all occasions" (Montaigne 2003: 294). Montaigne may have intended this as an expression of what is sometimes called the classical conception of virtue (Annas 2007). On that conception, virtues are dispositions to respond in certain ways to given kinds of situation. They are sometimes characterized as 'reliable' or 'robust' dispositions, but that seems to be a matter of emphasis. To say that courage is a disposition for valiant actions and emotions, on this view, just is to say that a courageous person is *reliably* valiant: that is, they show valiant behaviour whenever (or almost whenever) a situation affords or requires it. A familiar challenge to the classical conception comes from work in social psychology that allegedly supports a 'situationist' approach to action explanation. The central claim here is that an adequate explanation of our ethical or unethical behaviour makes no reference to character: our actions are supposed to be fully intelligible in the light of features of the situation we are placed in, showing character to be either epiphenomenal or even nonexistent (see Miller 2020). We want to set the situationist challenge to one side here. The view we are interested in is not that character plays no role in action explanation but that, consistently with acknowledging its explanatory role, we should resist a monolithic conception of character.

Montaigne himself is an eloquent advocate of that view:

> All contradictions can be found in me by some twist and in some fashion. Bashful, insolent; chaste, lascivious; talkative, taciturn; tough, delicate; clever, stupid; surly, affable; lying, truthful; learned, ignorant; liberal, miserly and prodigal: all this I see in myself to some extent according to how I turn, and whoever studies himself really attentively finds in himself, yes, even in his judgment, this gyration and discord (Montaigne 2003: 294).

There are two ways in which Montaignian character traits deviate from the classical conception. First, they are more fine-grained, or more context-dependent, than the classical conception allows. Montaigne observes that the same man "may be charging into the breach with brave assurance" while "later tormenting himself, like a woman, over the loss of a lawsuit" (294). That is to say, someone may have a disposition to behave valiantly *in a subset* of the situations that afford or call for valiant behaviour; they may have a disposition for (roughly speaking) cowardice in another such subset. Second, Montaigne denies that our ethical dispositions are consistent over time: "I give my soul now one face, now another, according to which direction I turn in" (293-94).

Now, if we conceive of virtues on the model of dispositional properties such as fragility or solubility, Montaigne's emendation of the classical conception will look puzzling. What could be the explanatory value of dispositions that are neither robust nor stable? What would be gained by describing an action as the exercise of a virtue conceived not only as highly context-dependent but also as fickle?

In turn, what could be the moral significance of the question whether a generous act was informed by Montaigne-style generosity (in effect, it might be said, the disposition to act generously, unless one doesn't)?

These are good questions, but they have, so we want to suggest, good answers. Put in general terms, our suggestion is that lack of robustness and stability does not have to make a virtue erratic or unintelligible. Montaignian virtues and vices come with their own distinctive sort of intelligibility. Commonsense psychology has rich resources to enable us to make such traits appear less erratic or irrational than they may initially seem. Admittedly, these resources are limited, and they often fail to secure full transparency. That, however, is no objection to the thesis that Montaignian character traits figure in our ordinary explanatory and evaluative practice. As Montaigne would be the first to agree: we are not fully transparent to each other, or to ourselves.

The starting point for developing this suggestion is a basic and familiar difference between properties such as fragility and properties such as generosity: viz. the latter involve a sensitivity to normative reasons. As Annas writes (expounding the classical conception of virtues): "A virtue, unlike a mere habit, is a disposition to act for reasons, and so a disposition that is exercised through the agent's practical reasoning; it is built up by making choices and exercised in the making of further choices" (2007: 516). More specifically, it has been suggested that to have a specific virtue is to be someone for whom certain kinds of facts count as reasons to do certain things (Schueler 2003: 81) or someone who is "sensitive" to relevant kinds of facts "as reasons for acting in certain ways" (McDowell 1998: 53). As a consequence, someone's virtues and vices may be said to reflect a person's values. The connection raises some delicate issues. Annas writes that "(t)o qualify as a virtue, a character trait must embody a commitment to some ethical value" (2007: 519). It would be a mistake, however, to equate generosity with a commitment to the value of generous behaviour, at least if such a commitment in turn is explicated as possession of a certain evaluative *belief*. A virtue is not a propositional attitude. We will use Gary Watson's notion of a person's 'evaluative orientation' to gesture towards the hard-to-articulate sense in which a virtue 'embodies' some ethical value (Watson 2004). To say that a person is generous is to say that they are apt to recognize, say, facts regarding others' needs or well-being as reasons for acting in certain ways, and that they are disposed to be responsive to such reasons. They have a character trait that amounts to taking up a certain position on what sorts of facts count as good reasons for action.[3]

A particularly helpful feature of Watson's notion is that it draws attention to the fact that no particular virtue *exhausts* an individual's evaluative orientation.

[3] An important question we cannot take up here is whether thinking of virtues and vices in this way should lead us to resist a dispositional account of character traits, or, instead, to insist on the distinctive nature of the relevant dispositions. A good starting point for consideration of that question would be the following passage from Nomy Arpaly's *Unprincipled Virtue*: "Why should Aristotle, or anyone else, believe that the praiseworthiness of an individual action depends on the character from which it stems? If one thinks of character as a stable disposition of some sort, the idea may seem strange. [...] The answer is that the mere frequency or predictability of an action does not matter at all to the moral worth of the actor, but these things may be signs of something relevant: *deep moral concern*. The pathologically fearless man or the well-trained soldier may have just as stable a disposition as the brave man to defend his city, but fearless or merely well-drilled actions do not express courage" (Arpaly 2003: 239).

No-one is simply or exclusively a generous person. You might be a person who is generous, honest and open-minded, among other things, whereas I may be honest, stingy and grumpy, and someone else may be generous, mildly corrupt, and open-minded. The various elements of someone's profile of virtues and vices inevitably (and intelligibly) affect each other. Someone who is generous and puritanically high-minded will have a different overall evaluative orientation—will be sensitive to different sorts of reason-giving facts—from someone who is generous and has imbibed a portion of what is sometimes called 'amoral familism'.[4] The way in which someone's generosity is embedded in their wider evaluative orientation will have implications for the range of situations in which they exercise their generosity. That a generous person fails to act generously in a situation that calls for generous behaviour does not necessarily mean their generosity is fickle or erratic. It may be intelligibly circumscribed or curbed by other virtues (or vices).

This provides the beginnings of a response to the charge of opacity directed against Montaignian virtues. What may initially look like an erratically context-dependent exercise of generosity may, on closer acquaintance, turn out to be intelligible in the light of the agent's wider evaluative orientation. The response can be further developed by noting another distinctive feature of the sort of explanation in which virtue terms pull their weight. There are two perspectives on someone's reasons that are relevant in the context of reason-giving explanations: the agent's own perspective and the interpreter's perspective. The agent's perspective, of course, is paramount. In trying to make sense of someone's intentional actions we must surely be interested in *their* conception of their reasons—in the considerations in the light of which they are acting. But our own view of what they have reason to do can affect our interpretation in a number of ways. Something that may strike us as a rationally unintelligible feature of their behaviour may in fact reflect a disagreement over their reasons.[5] The impression that someone's exercise of a certain virtue is erratic may be a case in point. We may find their reluctance to exercise a certain virtue in a situation which, we are convinced, calls for its exercise hard to understand, given that they *seem* to manifest the virtue in other kinds of situation. But the puzzle, of course, reflects our perception of the situation. Perhaps from the point of view of the agent's evaluative orientation, the two kinds of situations are relevantly different: one of them calls for the exercise of generosity, the other, say, for the exercise of justice. If we continue to think that their perception is mistaken, there will be work to be done for us in trying to understand their (as we see it, flawed) outlook. But it is not that we are confronted with a capricious disposition. Consider also the myriad ways in which commonsense psychology attempts to understand examples of apparent instability in someone's character traits. As Montaigne famously observed, "(t)he mayor and Montaigne have always been two, with a very clear separation" (2003: 941). Our professions or social roles may impose elements of an evaluative orientation on

[4] This raises familiar questions regarding the unity of the virtues. When operating in conjunction with 'amoral familism', it might be said, generosity is not (in Philippa Foot's phrase) "operating as a virtue" (Foot 2002: 16).

[5] Compare McDowell observation that "(f)inding an action or propositional attitude intelligible, after initial difficulty, may not only involve managing to articulate for oneself some hitherto merely implicit aspect of one's conception of rationality, but actually involve becoming convinced that one's conception of rationality needed correcting, so as to make room for this novel way of being intelligible" (1998: 332).

us from which, in our better moments, we manage to distance ourselves: "For all of being a lawyer or financier, we must not ignore the knavery there is in such callings" (ibid.). Again, there are familiar narrative structures that appear to enable us to make sense of the evolution of someone's character, as when an academic famed for his sharp tongue mellows into an avuncular figure. Finally, consider one way Montaigne himself appears to make sense of what he describes as the multiple 'contradictions' in his character: "irresolution seems to me the most common and apparent defect of our nature" (2003: 290). Whether or not it is the most common defect, it seems right that an apparent inconsistency in someone's evaluative orientation may reflect a genuine ambivalence.

Let us return to the single instance intuition. Suppose our ordinary conception of the virtues is not monolithic but allows for the various—complicated, but often intelligible—sorts of instability and context-dependence Montaigne highlights. Then the fact that a single generous act provides no adequate evidence of a *firm and unchangeable* disposition of generosity cannot be used to put pressure on the Dependence thesis, or on the idea that our ordinary practices of holding each other responsible are in keeping with the Dependence thesis. For the Dependence thesis may now be developed like this: in acting generously, a person shows themselves to be generous, in the sense that there is some generosity in them or, as we might say, they 'can be' generous. In effect, abandoning the monolithic conception of character amounts to lowering the requirements for the global use of virtue terms. That I am not always acting generously in any conceivable situation affording it does not mean that I am not generous or that my current action should not be interpreted as manifesting generosity. To lower the requirements is not to emasculate them, though. Advocates of the Dependence thesis are committed to the view that only if a given virtue term finds a foothold in an agent's evaluative orientation will it be appropriate to apply that term to a particular action of theirs. We want to suggest, though, that the single instance intuition, on careful consideration, does not challenge that commitment.

Recall the dog-kicker in Hurka's example. It seems right that when we see someone 'kick a dog from an evident desire to hurt the dog just for the pleasure of doing so', we would be inclined to judge the act to be vicious, even if we have no evidence of a firm disposition of vicious behaviour on the part of the agent. But would we take that local judgement to be wholly independent of questions about the agent's character? Consider the following variation on the story. Suppose we know the man who is kicking the dog, or at least think we know him; specifically, we think we know there is not a smidgen of viciousness in his character. On Hurka's account of our 'everyday understanding of virtue', that should not affect our local judgement. We should take our (presumed) background knowledge to be simply irrelevant when it comes to our judgement that he acted viciously. That seems implausible. A more lifelike description, surely, is that we would be puzzled and, at least initially, unsure what to think. Various kinds of questions would arise: was our impression correct that his act was intended to hurt (or did he perhaps feel threatened by the dog)? Was he in a normal state of mind? Do we know him as well as we think we do—or does his action possibly bring to light some hitherto hidden or repressed facet of his character? Were we wrong to take him to be a stranger to viciousness? That we should feel compelled to ask such questions suggests that we do not, as Hurka's interpretation of our 'everyday understanding of virtue' would suggest, take the local use of a virtue term to be wholly detached from a global use. It is not that the local use commits

us to the claim that the dock-kicking must have issued from a 'firm and unchangeable character trait'. Still, it would normally be taken to be somewhat revealing of what sort of person the dog-kicker is. That is why we would tend to be puzzled: pending answers to our questions, we would do well to suspend judgement as to how his action is to be understood and assessed.

How about the intuition that someone may coherently be described as having 'uncharacteristically' done 'something quite heroic'? If 'uncharacteristic' means that the act did not issue from a stable disposition, the point does seem intuitive but it is compatible with the Dependence thesis (on a non-monolithic conception of character). If 'uncharacteristic' means that the act tells us nothing whatsoever about what sort of person the agent is, the intuition arguably wanes. Recall the award for bravery. Even if the award was in recognition of a single heroic exploit, and even if the soldier had hitherto not shown much of a disposition for courageousness, we (and he) would tend to think that his valiant act revealed *something* about who he was (perhaps a recently acquired, and not wholly robust, streak of bravery).[6]

To summarize, we have tried to defend the Dependence thesis against the charge of revisionism, by suggesting that that charge is predicated on an implausible account of our ordinary conception of character. There is, we grant, an element of revisionism in the classical conception of the virtues. What is revisionary, however, is not the Dependence thesis but the monolithic conception of character—something critics of the thesis tend to grant. We now want to turn to the second lesson that has been drawn from the single instance intuition: the charge that the Dependence thesis lacks an intelligible rationale.

4. The Role of Character in Reason-Giving Explanations

As we saw, a natural way to press the question of the rationale is to ask why we should give preferential treatment, as it were, to one of two apparently identical acts. What makes a generous act that manifests a character trait of generosity better—more deserving of moral approbation—than a generous act that does not? It is agreed on all hands that the moral merit of an action turns on the agent's motives or, as Hurka puts it, on their "occurrent motivation" (2006: 70). But to ascertain whether an act was genuinely generous, rather than, say, actuated by the desire to impress others, it may seem, we only need to look at the agent's current attitudes—notably their beliefs and desires, and their role in leading the agent to act. We only need to consider their "current motives, apart from any

[6] But can we be justified in calling a person generous who keeps performing ungenerous acts? Is there a minimal condition that needs to be satisfied to warrant attributions of virtue, as conceived by the non-monolithic view? The question deserves more extensive discussion than we can offer here, but we would like to make two points. First, we would suggest that the idea of a 'minimal condition' may best be spelled out not in terms of a statistically relevant incidence of (e.g.) generous acts, but in terms of the demand for an account of how it is that a putatively generous person keeps performing ungenerous acts. As we illustrated earlier, commonsense psychology has a range of relevant resources at its disposal. Second, in practice it will often be hard to know whether some such account is available. Thus, we should (once again) heed Montaigne's advice: "a sound intellect will refuse to judge men simply by their outward actions; we must probe the inside and discover what springs set me in motion. But since this is an arduous and hazardous undertaking, I wish fewer people would meddle with it" (2003: 296).

connection to longer-lasting traits", traits that amount to "external features" of their act (Hurka 2006: 71).

Once again, the right response to this challenge, we suggest, is to probe the terms in which it is framed. If occurrent motivation is pitted against longer-lasting traits, it looks puzzling why the latter should matter to moral judgements. The question we want to press in response is whether occurrent motivation can generally be understood *in isolation* of aspects of the agent's 'evaluative orientation' and so of their character. Hurka does not argue for an affirmative answer to this question; he simply takes that answer for granted. We want to suggest that there is a case to be made for a negative answer, and that if correct, that argument would deliver a compelling rationale for the Dependence thesis. The argument we have in mind can be extracted from Fred Schueler's work on what he calls teleological explanations of actions. What he means by this is the utterly familiar sort of explanations we use when we make sense of our own and others' intentional actions as "inherently purposive" (Schueler 2003: 1). Central to such explanations are the considerations the agent takes to provide them with reasons for action, i.e. considerations that support or justify or count in favour of acting in a certain way. In Schueler's discussion, character emerges as the solution, or part of a solution, to a puzzle over the explanatory force of appeal to the agent's reasons. We briefly set out Schueler's suggestion, and then consider how it bears on our understanding of 'local' uses of virtue terms.

Here is the puzzle. We often have reasons for and against a certain course of action. Suppose you accept a job offer, and we explain your decision by reference to the relevant reasons. But you might have refused the offer, in which case we would have explained your decision by reference to the opposing reasons. Thomas Nagel uses this example to illustrate a completely general concern about reason-giving explanation: it can seem puzzling how such explanations can be genuinely illuminating. Nagel puts the matter like this:

> Intentional explanations, if there is such a thing, can explain either choice in terms of the appropriate reasons, since either choice would have been intelligible if it occurred. But for this very reason it cannot explain why the person accepted the job for the reasons in favor instead of refusing it for the reasons against (Nagel 1986: 116).

If either action is open to an equally illuminating explanation, we seem to lack an account of why the person accepted the job *rather than refusing it*. Now, it seems clear that in some cases, Nagel's worry is easy to dispel. The reasons against may be so obviously flawed or at least obviously less weighty that any remotely rational agent will recognize the greater force of the reasons in favour. But Nagel is surely right that not all situations are like that. Either decision may seem rational, and it may look as if an explanation of the person's accepting the offer in terms of the reasons in favour only appears illuminating so long as we do not ask 'Why did they not instead refuse it for the reasons against?'

Schueler's move is to suggest that our ordinary practice of reason-giving explanation has richer resources than Nagel allows. We may find the person's decision to accept the job intelligible in the light of the sort of person they are. Character traits are an important ingredient of such a conception, and they may bear on the sort of case Nagel highlights. Even in a scenario in which there are equally respectable reasons in favour and against, it may be the case that no-one who

knows the person will be surprised that they accepted the offer. Perhaps one of the respectable reasons for taking the job is that the job comes with a higher salary, though this, we may suppose, is counterbalanced by a higher teaching load. In view of a mildly avaricious streak in their outlook, it may come as no surprise that they were unable to resist the offer. The example is banal, but the proposal it illustrates seems suggestive. If being avaricious means, in part, being someone "for whom certain kinds of facts count as reasons to do certain things"—or, significantly, count as reasons "of a certain strength" (Schueler 2003: 81)—then someone's being avaricious will be precisely the sort of thing that can make it intelligible which of two finely balanced sets of reasons carries more weight with them. In this way, appeal to someone's character traits can play a crucial role in understanding their 'occurrent motivation'. Making sense of someone's action in the light of thei reasons may call for reflection on 'the sort of person they are, insofar as this sheds light on how it is that they are responsive, or gives a certain weight, to some reasons and not to others.

As Schueler remarks, this proposal about the explanatory role of character has an interesting bearing on how we should think about the nature of character. In particular, it would suggest that there is a certain explanatory depth to the way character traits help to make intentional actions intelligible, which would seem to count against a 'purely dispositional' view of character (or at least against the idea that our ordinary conception of character is adequately characterized by a purely dispositional view) (see Schueler 2003: 80f). We cannot pursue these important issues here, nor can we address the question whether Schueler's point about the explanatory role of character should be seen to hold as a matter of complete generality, or merely in certain special contexts (such as the ones highlighted by Nagel's puzzle). For current purposes, we can confine ourselves to two observations.

One is that if Schueler's proposal is on the right lines, the Dependence thesis can be seen to be rooted in our ordinary explanatory practice. The question of how an action reflects on the agent's character matters in the context of our practice of *evaluating* the action because it matters in the context of *understanding* what they are doing and why they are doing it—for example, whether they are doing the right thing for the right reasons. The baking analogy we drew earlier (would the Dependence thesis not encourage a differential assessment of actions that would be unfair and unmotivated in a way akin to awarding a lesser prize to a cake on the grounds that it was produced by a novice baker?) is flawed in just the way Aristotle tells us that analogies between the virtues and the arts tend to be flawed. While "the products of the arts have their goodness in themselves" (1980: 1105^{a}), the goodness of an action depends on its motivation, which (often and possibly invariably) can only be adequately understood in the light of the agent's 'evaluative orientation' and so their character.

Our second observation is that while Schueler's move demands that character traits can play a substantive explanatory role, it does not require a monolithic conception of character. We may make sense of a generous action in the light of the agent's having a generous streak or our sense that they are someone who 'can be' (in certain, perhaps hard to codify, contexts) generous. What matters is not whether they have a stable disposition but whether their responsiveness to the reason for their action manifests some, however fragile and possibly short-lived, aspect of their 'evaluative orientation'. As indicated earlier, there are a range of ways in which the context-dependence and instability of such traits may be

rendered intelligible, though it is true that in this enterprise we more or less quickly come up against certain limits of intelligibility. It is not clear, though, that this counts against Schueler's proposal. As Strawson remarked: persons "may puzzle us at times" but that "is part of [...] reacting" to them as to persons (Strawson 1985: 21).

There is a certain irony in the dialectical position we have reached. On Hurka's view, the Dependence thesis amounts to a revisionary philosophical theory, since ordinary assessments of actions are centred on questions about the 'occurrent motivation' rather than 'long-lasting traits'. If Schueler is right, this diagnosis is itself premised on a revisionary view, viz. a 'belief-desire model' of action explanation that ignores or distorts the explanatory role we ordinarily assign to the agent's perception of their normative reasons for action. There is also, however, an important point of agreement with Hurka's approach. We should be clear about the distinction between 'descriptive' vs 'revisionary' ethics (to adapt Strawson's well-known distinction between two styles of doing metaphysics) (Strawson 1959), and virtue ethics had better be alive to the possibility that some of its traditional tenets may be revisionary.

5. Conclusion

In sections 3 and 4 we have outlined a version of the Dependence thesis that is not committed to the monolithic conception of character. The upshot is that understanding, explaining, and evaluating an action requires a reference to the character of the agent, which, however, need not be articulated in terms of fully stable and generalisable dispositions. But is it not true, someone might say, that evaluating someone's character—however we understand it—is precisely a matter of looking at what they *do*? And if so, how could the former ever impose a requirement on the latter?

Though compelling, we maintain that these questions do not pose a challenge to our main thesis. On the contrary, they help us clarify the nature of the Dependence thesis. For insisting that the 'local' use of virtue terms cannot be independent of their 'global' use does not commit us to the stronger claim that there is only one correct direction of explanation here, going from persons to actions. Rather, our thesis is compatible with the converse claim that an evaluation of someone's character cannot prescind from their conduct. This suggests an interesting diagnosis: perhaps we encounter difficulties in pigeonholing the way in which we use virtue terms in *either* their global *or* local usage precisely because this distinction is somewhat artificial to begin with. To assess whether a virtue is instantiated, we may need to consider actions and agents together. And once again, far from being revisionary, this seems to be in keeping with the way in which we ordinarily apply virtue terms.

A similar point is made by Kieran Setiya in *Reason without Rationalism*. Rather than discussing whether virtue terms apply primarily to actions or persons, he considers whether right action should be explained in terms of ethical virtue or vice versa, but we take the two questions to have a sufficiently similar structure. And Setiya concludes:

> [...] although I am arguing for a metaphysical connection between ethical virtue and practical reason, I do not claim that the connection is *asymmetric* in any interesting way. We can say what it is to be a reason for action in terms of ethical virtue,

> or so I will claim. But that is not to say that the virtues of character have explanatory primacy. The connection between reason and virtue runs in both directions: it is a matter of reciprocity, not priority (Setiya 2007: 5).

Similarly, we want to suggest that the connection between the global and the local use of virtue terms runs in both directions. Evaluating agents and actions is a matter of reciprocity rather than explanatory (or metaphysical) priority. And in effect it is not clear why it should be that *either* someone is generous because (independently) their actions is generous *or* their action is generous because (independently) they are generous. Drawing such a sharp distinction might only create an unnecessary ravine that, in fact, we need not bridge.

References

Annas, J. 2007, "Virtue Ethics", in Copp, D. (ed.), *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press.

Aristotle 1980, *The Nicomachean Ethics*, tr. by D. Ross, J.L. Ackrill and J.O. Urmson, Oxford: Oxford University Press.

Arpaly, N. 2003, *Unprincipled Virtue: An Inquiry into Moral Agency*, Oxford: Oxford University Press.

Foot, P. 2002, "Virtues and Vices", in her *Virtues and Vices*, Oxford: Oxford University Press, 1-18.

Hurka, T. 2006, "Virtuous Act, Virtuous Dispositions", *Analysis*, 66, 1, 69-76.

Hursthouse, R. 1999, *On Virtue Ethics*, Oxford: Oxford University Press.

McCormick, M. and Schleifer, M. 2006, "Responsibility for Beliefs and Emotions", *Paideusis*, 15, 1, 75-85.

McDowell, J. 1998, "Functionalism and Anomalous Monism", in his *Mind, Value, Reality*, Oxford: Oxford University Press, 325-40.

Miller C. 2020, "Empirical Approaches to Moral Character", *Stanford Encyclopedia of Philosophy*.

Montaigne M. de 2003, *The Complete Works*, tr. by D. Frame, London: Everyman's Library.

Nagel, T. 1986. *The View from Nowhere*, Oxford: Oxford University Press.

Schueler, F. 2003, *Reasons and Purposes*, Oxford: Oxford University Press.

Setiya, K. 2007, *Reason without Rationalism*, Princeton: Princeton University Press.

Strawson, P.F. 1959, *Individuals: An Essay in Descriptive Metaphysics*, London and New York: Routledge.

Strawson, P.F. 1985, *Skepticism and Naturalism*, London: Methuen.

Watson, G. 2004, "Two Faces of Responsibility", in his *Agency and Answerability*, Oxford: Clarendon Press, 260-88.

Book Reviews

Queloz, Matthieu, *The Practical Origins of Ideas: Genealogy as Conceptual Reverse-Engineering*.
Oxford: Oxford University Press, 2021, pp. xiv + 304.

A brand-new genealogical season seems to be starting, but despite this growing popularity, there is still a lack of common ground on what genealogy is and what it stands for, and an alarmingly vast variety of conceptions is still available on the market. Moreover, being in the middle of the notorious analytic-continental divide, cultural as well as philosophical misunderstandings abound.

Before Matthieu Queloz, no recent author had ever addressed the question of genealogy as a philosophical method in such detail. The interest aroused by *The Practical Origin of Ideas* is not unexpected, since it makes available a well-conceived conception of philosophical genealogy, whose perspectives and meta-philosophical ambitions are clear and defined, though open-ended and plural. Moreover, his work taps into the manifold of genealogical conceptions in circulation, both, and most evidently, from genealogies traceable to the influence of Bernard Williams,[1] as well as those inspired by the Foucauldian tradition.[2] The author undertakes two distinct but closely related operations: the methodological exposition of what he calls "pragmatic genealogy" and the rediscovery of a hitherto ignored historical tradition of this method. We thus realize that great authors of the past such as Hume and Nietzsche, and more recently Edward Craig, Bernard Williams and Miranda Fricker, can be plausibly assigned to this tradition. The two projects are mutually enlightening: through the presentation of the method, it is possible to bring out instances of it, which in turn allows us to test its qualities (18-19).

The book's first three chapters are devoted to laying out pragmatic genealogy's theoretical framework and methodological assumptions. The first chapter moves from some questions and suspicions concerning our most abstract ideas. We inherit venerable ideas, such as those of truth, justice and knowledge, the practical purpose of which is often unclear to us; nevertheless, our actions as individuals and as human communities are guided by these same ideas. The method of pragmatic genealogy allows us to reveal what such ideas do for us, a result we achieve through the production of a peculiar historical-philosophical artefact: a rational and historical (sociological, psychological) narrative that explores how we have developed them. To be more precise, the *explananda* of this method are *conceptual practices*, that is, practices "[…] essentially shaped by sensitivity to conceptual norms or reasons—take away the idea in terms of which those norms and reasons are articulated, and the practice collapses" (3).

I will try to summarize the nature of this method for the benefit of the rest of the review. The outcome of pragmatic genealogy should not be thought of as a succession of historical facts, but as a model analogous to those in science,[3] which provides us with a perspicuous view of the conceptual practice examined. The

[1] Fricker, M. 2007, *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford: Oxford University Press.
[2] Cf. Koopman, C. 2013, *Genealogy as Critique: Foucault and the Problems of Modernity*. Bloomington: Indiana University Press, for more on these two different conceptions.
[3] An idea that has precedents: Cf. Kusch, M. 2009, "Testimony and the Value of Knowledge", in Haddock, A., Millar, A. and Pritchard, D. (eds.), *Epistemic Value*, Oxford: Oxford University Press.

Argumenta 9, 1 (2023): 209-222
ISSN 2465-2334
© 2023 The Authors
DOI 10.14275/2465-2334/202317.boo

model in question is dynamic, representing a changing object, and emerges in two stages: a *fictionalizing* stage and a *historicizing* stage. The former requires "a maximally ahistorical setting",[4] which Craig, from whom Queloz draws inspiration, suggestively calls the "state of nature".[5] In this setting, we represent the traits of a conceptual practice (a proto-practice) whose function corresponds to the most basic conceivable function of the conceptual practice we intend to explain. By gradually increasing the complexity of the factors involved in this toy-society, it is possible to observe, step by step, the modification of the practice in response to ever-changing needs. In this way, it is possible to break down, analyse, compare and, above all, present in sequence those instrumental relations inherent in conceptual practices that in real life we can observe only synchronically. In the second stage, we move from an ideal model to a model based on the actual history of a human community: the conceptual practice under investigation is thus historicized. It is a matter of incorporating historical needs and pressures into our model and showing how that practice changes in response to them. If in the first stage, it is possible to detect the *practical needs* underlying the practice under scrutiny, the second stage shows us the *historical contingencies* that shaped the proto-practice into what it is today.

According to the author, this explanatory procedure, besides being a clear example of philosophy as model building, is analogous to a reverse engineering operation. With clear reference to the influential philosophical project known as conceptual engineering, he calls pragmatic genealogy an instance of *reverse conceptual engineering*. The second chapter is thus devoted to the presentation of seven virtues of reverse conceptual engineering, to which are added three distinctive benefits of pragmatic genealogy as a form of conceptual engineering: explanation without reduction, normative significance, and the facilitation of responsible conceptual engineering. In the next chapter, Queloz proceeds to examine the strengths of his favoured method, as compared to other forms of reverse conceptual engineering (above all the *paradigm-based explanation*). In particular, he identifies two kinds of conceptual practices that would be hardly analysable without pragmatic genealogy: *self-effacingly functional* practices and *historically inflected* practices. The former has a rather elusive functional requirement: for the practice to be properly functional, the agents must not have access to its function when they engage in it. The latter are those current practices in which the link to the basic needs they were serving when they arose has not been conserved.

The most hermeneutically inspired chapters, in which the author aims to bring to light the hidden philosophical tradition of pragmatic genealogy, are devoted to Hume and Nietzsche. Queloz comes to Hume's aid, in the fourth chapter, defending him from the accusation of producing a merely conjectural form of history. Similarly, in the fifth chapter, he refutes the charge addressed to Nietzsche, who allegedly traced a scarcely documented historical genealogy: Queloz shows how to correctly understand their purposes through the lens of pragmatic genealogy. In addition to the exegetical insights contained in these chapters, Queloz highlights some peculiarities of his method by drawing on the genealogies of the two authors reviewed; the possibility of vindicatory genealogy is exempli-

[4] Cf. Fricker 2007: 108-109.
[5] Craig, E. 1990, *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*, Oxford: Oxford University Press.

fied by Hume's treatment of the virtue of justice, while Nietzsche's work is summoned in support of the possibility of employing this method to avoid what the German thinker thought was the philosophers' ancient defect of thinking ahistorically.

The chapters concerning Craig, Williams and Fricker (§6, §7, and §8, respectively) allow Queloz to substantiate his historical thesis and showcase further benefits of his proposal. As Williams once warned, "the state of nature is not the Pleistocene",[6] and the chapter on Craig further clarifies what the state of nature is and what its implications are. Queloz takes the occasion to argue in favour of the compatibility between Craig's approach, incorporated in his method, and the principles of factivity and non-analysability of knowledge of the widespread *Knowledge first* epistemological conception.

The chapter on Williams is in my view pivotal to this book. *Truth and Truthfulness* is still considered a significant work today; despite this, the aspects that were most important to the author in writing this book are rarely considered. Queloz offers, perhaps for the first time, a well-documented clarification and a strenuous defence of the author's intent and method. We find here one of the most representative examples of pragmatic genealogy, one that is not only extensive but also paradigmatic, since it is the genealogy of a self-effacing practice, the best-suited field of action of this method. From Williams we learn how an exclusively instrumental use of practices related to truthfulness along with access to the function of these would profoundly destabilize them to the point of collapse: if all individuals expected truthfulness in the practices of others while reserving for themselves the possibility of not being truthful for their own benefit, this would soon result in the collapse of these practices: we could not expect truthfulness from anyone. Concealing their own function is vital for these practices to remain stable. As a result, Williams' vindicatory genealogy leads us to an apparently controversial result: it shows how it is possible to value intrinsically a conceptual practice based on an abstract and venerable concept while at the same time continuing to value this practice instrumentally. In support of Williams, Queloz defends the compatibility between valuing a conceptual practice intrinsically and valuing it instrumentally.

Chapter eight is devoted to the genealogy contained in Miranda Fricker's influential book *Epistemic Injustice*. Here, Queloz has a chance to show that pragmatic genealogy can be proposed as an ameliorative project of our practices and not merely as a descriptive survey. Taking an ameliorative outlook is, in a few words, about trying to change our current practice to what we believe it should be. This perspective has attracted great interest within conceptual engineering, devoted among other things precisely to exploring the possibility of modifying our representational devices, such as concepts. However, pragmatic genealogists can also pursue an ameliorative approach, as exemplified in Fricker's work. Indeed, the retro-engineering of conceptual practice allows us to identify the developments that resulted in a practice that we believe is not the best possible. As Fricker's genealogy clearly shows, this opens the way for an ameliorative process. She brings in a political dimension precisely at the exit from the State of Nature: it consists of the creation of social groups and the consequent phenomena of social categorization. Then, she shows how the testimonial injustice that still abounds

[6] Williams, B. 2002, *Truth and Truthfulness: An Essay in Genealogy*, Princeton: Princeton University Press, 27.

today is the result of pressures opposed to a virtuous division of epistemic labour, inviting us to cultivate the virtue of testimonial justice.

Having finished his close examination of the work of past genealogists, Queloz turns back to the exposition of his method. In the ninth chapter, the normative ambitions of pragmatic genealogy are defended: Queloz presents and responds to four increasingly specific objections that sum up the most common criticisms addressed to normatively ambitious genealogical explanations. First, the charge of genetic fallacy is dismissed. Queloz presents two different forms of the genetic fallacy. The former cannot threaten his method; the latter is committed only by inferring something about the justification of a conceptual practice from irrelevant information about its formation process, which is entirely avoidable in a pragmatic genealogy. He then proceeds to describe two kinds of conceptual practices in which the formation process carries normative weight. The second charge, which focuses on lack of continuity, is avoided altogether, probably because it applies only to far more traditional genealogies. Queloz's genealogy does not assume that there must be continuity between the conditions under which a conceptual practice arose and those that survive today, but on the contrary, is designed to reveal it. He also rejects the claim that pragmatic genealogy can only deal with practices that emerged in connection with anthropological universals, which would severely narrow its scope: he shows in some detail how pragmatic genealogy can also deal with extremely local and contingent practices. The last objection, which points the finger at the arbitrariness in the attribution of needs on which Queloz's method is based, is partly overturned and partly accepted. This method makes it possible to account for the attributions of needs since these must be systematically traced back to basic and increasingly less contestable needs. However, a central role is indeed played by the genealogist's point of view, but this is a welcomed aspect of this method, which does not assume that there is an extra-subjective point of view for such matters.

The last chapter is spent on some meta-philosophical considerations. Here two possible approaches that pragmatic genealogy can encourage are introduced: a Socratic inquiry grounded in pragmatic inquiry and the practice of philosophy as a humanistic discipline. The latter approach, evidently Williamsian, reveals how good genealogical practice requires the maximum integration of insights gained from the other humanistic disciplines and in the social sciences, abandoning the idea of a pure philosophical inquiry independent of other forms of knowledge.

The Practical Origins of Ideas reflects Queloz's erudition and his well-rounded knowledge of the field of inquiry, as well as his remarkable clarity and care in exposition. As already mentioned, Queloz brings new life to the thought of the late Bernard Williams, an author of whom he is an eager connoisseur given how confidently he masters his vast and various philosophical production

The weaknesses of this book are mainly architectural. The author has made a very hard but understandable choice in the economy of the text, by not producing *ad hoc* instances, choosing instead to exploit past genealogies that he has reread (or rediscovered) as pragmatic genealogies. This constitutes a burdensome constraint because it does not allow the choice of more didactic examples and ties their exposition to previous exegetical passages.

In addition, the historical thesis regarding the tradition of pragmatic genealogy, although presented with abundant interpretative suggestions, is not treated

in a sufficiently extensive and systematic manner, which might give the impression that it is ultimately not of primary importance. If, as we have said, the historical thesis constitutes one of the two levels upon which this book is developed, it is surprising how no general chapter has been devoted to the alleged philosophical tradition of pragmatic genealogy; where to discuss, for instance, the reasons why this has remained unseen through the years. Instead, we are faced with a series of chapters in which, individually, methodological similarities of varying strength are detected, but whose overall historical nexus remains elusive to us. Tracing similarities a posteriori in the light of a given systematization is not in itself illicit, but it is not sufficient on its own to constitute a historical account.

Another theme presented in several places but partially unaddressed is conceptual engineering. Conceptual engineering is explicitly referred to by the author in several places (17, 30, 193, 208), and of course, it is integral to one of the book's main themes: reverse conceptual engineering. In light of this, we would expect a close exploration of the relation between these two philosophical enterprises throughout the book. Unfortunately, we must settle for a few rather general passages, such as the one about how pragmatic genealogy encourages responsible conceptual engineering (41). In the absence of a detailed examination of the methodological assumption of these two projects, it is not even clear whether they are compatible and integrable.

In any case, Queloz's book is still a vigorous attempt to undertake a methodological and rigorous approach to genealogy, an effort that appears to be decidedly well-directed and capable of yielding valuable results. We now have only to look forward to developments in a methodological direction and an applicative one.

Independent researcher FRANCESCO ALBENZIO

Lieto, Antonio, *Cognitive Design for Artificial Minds*.
New York: Routledge, 2021, pp. xiv + 119.

The collaboration between artificial intelligence (AI) and cognitive science is a long-lasting debated topic and it is very deeply intertwined with the theoretical foundations of these two disciplines. Even though AI and cognitive science are different fields, with different aims, methods, and applied results, they share at least two things, speaking from a very wide perspective: 1) the object of research: intelligence and cognition; 2) a general interdisciplinary and transdisciplinary approach. If for some respects the former claim is correct, and therefore intelligence and cognition can be considered as two partially overlapping notions, the latter is a sort of necessary condition for the birth of both: AI in the mid-twentieth century and cognitive science a couple of decades later. Nevertheless, it was through interdisciplinarity that these two fields could give rise to a common target, being AI from the very beginning dedicated to the simulation of "every aspect of learning and other features of intelligence"[1] and cognitive science to the study of thought

[1] From the Dartmouth proposal of 1955 and printed as McCarthy, J., Minsky, M.L., Rochester, N., & Shannon, C.E. 2006, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955, *AI Magazine*, 27, 4, 12, DOI: 10.1609/aimag.v27i4.1904

and mental phenomena by putting together aspects of psychology, philosophy, linguistics, neuroscience, anthropology, and computer science, especially AI.

One may wonder why AI should not be considered as a fully cognitive discipline, rather than an engineering and technological one, given that its aim is to simulate every feature of intelligence. This is related to the ambiguity of the notion of simulation. To simulate a performance of a task that is considered to require normally human intelligence is different from simulating the underlying mechanisms and processes enabling the intelligent behavior and the cognitive performance. Only in the latter sense the notion of simulation has been adopted by cognitive science and, in return, cognitive science has become (also) a computational discipline. The distinction between a more engineering approach and a more psychological one to AI is not new and is part of the evolution of the discipline since AI was mainly symbolic driven,[2] but the more recent approaches to AI has renewed the connection between AI and the study of principles, processes, and mechanisms upon which intelligence is based. Many of the new approaches are biologically and neurologically inspired, situated, evolutionary, dynamical, and embodied, so their biological plausibility is at the core of this new approach as much as in the new approaches to cognitive science.[3] Within this new framework Lieto speaks about a rebirth of a collaboration between AI and cognitive science, a collaboration that is grounded on the old ideas of simulation and computational modeling of cognitive capabilities.

The computational cognitive science that uses cognitive modeling involves some problems, among which the main one is the problem of model. What makes a computational model a cognitive one? What are the right and relevant constraints to build a model that is not merely a system producing the same performance in specific tasks as the humans do? As the author states, "'functional' systems (in the sense explained in the book) cannot be considered artificial models of cognition if they are not additionally equipped with 'structural constraints'" (93). This is effective if one wants to explain how mind and brain work (the main aim of the cognitive/psychological AI), but also if the overall goal is to achieve systems that are capable of a suitable interaction with human beings. It is not by chance that these issues are addressed especially in some recent AI trends, such as, for example, robotics (in particular, social robotics[4]), explainable AI, and artificial life.

Starting from these premises, the focus of Lieto's proposal is on cognitive architectures, a notion that was introduced by Newell in his attempt to define a unified theory of cognition.[5] They are abstract models between the high-level cognitive capabilities and their neural/bodily implementation, so they are at an intermediate level and their characterization as an integrated mechanism is what allows to build a computational counterpart of them in an artificial system. In

[2] See for example Winston, P. 1984, *Artificial Intelligence*, *2nd Edition*, Reading: Addison-Wesley.
[3] Cordeschi underlines the fact that new AI, with new models associated to the research projects of cybernetic period, is, in many cases and from this respect, the same as a new cognitive science. See Cordeschi, R. 2008, "Step Toward the Synthetic Method: Symbolic Information Processing and Self-Organizing Systems in Early Artificial Intelligence modeling", in Husbands P., O. Holland, and M. Wheeler (eds.), *The Mechanical Mind in History*, Cambridge, MA: MIT Press, 219-58.
[4] On this topic see Dumouchel, P. and L. Damiano 2017, *Living with Robots*, Cambridge, MA: Harvard University Press.
[5] Newell, A. 1990, *Unified Theories of Cognition*, Cambridge, MA: Harvard University Press.

other terms, a cognitive architecture is a model of one or more cognitive capabilities *and* its software implementation in a computational cognitive model. The more interesting cognitive architectures are, clearly, the more general ones, i.e. the ones modeling the cognitive capabilities at the highest degree of integration among intelligent features. The intermediate nature of cognitive architecture makes the problems of relevant constraints of modeling a crucial one to achieve an actual model of cognitive processes. In fact, the problem of right model is *the* problem of computational cognitive science using AI systems, as the assumption that the relevant constraints can be identified is the strongest one, from a methodological and epistemological point of view, to achieve both a "working" cognitive artificial systems and an explanation of the cognitive process.[6]

The cognitive architectures analyzed in the volume are probably the most well-known: SOAR and ACT-R,[7] starting from which many models have been developed in the last forty years. It is worth it to mention that they both started as symbolic architecture, but at least in the case of ACT-R many models developed within this general framework are hybrid, i.e. they mix symbolic and subsymbolic processes. One of the main features of many cognitive architectures is that they have a modular structure, which they derive from a well-established idea of mind that is typical of the classical, symbolic cognitive science and philosophy associated to it, especially by Fodor.[8] According to the modularity of mind view at least a part of cognition is carried out by modules, that is mental or neural structures with a specific function. Even though the modularity of a cognitive architecture is not strictly committed with modules that are characterized by the properties required by the theory, a modular structure is very well suitable to be described in a symbolic, discrete, and functional way, and in this way implemented in a software structure. For this reason, it appears to be even more convenient from a methodological point of view than from an epistemological one. A mechanistic integrated system is easily describable as a modular structure, which, in addition, fosters the possibility to build artificial systems with a hybrid way to process information, as it seems it should be the case. Or, at least, this is the view stated by Lieto.

The choice of SOAR and ACT-R is not by chance. They are two cognitive architectures in which knowledge representation is crucial and a very relevant part of the architecture. The knowledge level, to use a terminology by Newell, of both, however, is problematic for some respects, in particular for the limits that Lieto finds in "the limited size and the homogeneous typology of the encoded and processed knowledge" (65). If the former is roughly self-explanatory, the latter refers specifically to a semantic capability, i.e. the capability to categorize. Psychological research of the last fifty years has highlighted a big variety of this capacity even in the same cognitive agent, that is the human being. Heterogeneity means, therefore, flexibility, and the core of the author's proposal is a cognitive architecture

[6] And this is separate from the psychological and/or biological plausibility of the constraints. For a discussion on this see Cordeschi, R. 2002, *The Discovery of the Artificial. Behavior, Mind and Machines Before and Beyond Cybernetics*, Dordrecht: Kluwer.

[7] For a wide review of cognitive architectures see Samsonovich, A.V. 2010, "Toward a unified catalog of implemented cognitive architectures (review)", in Samsonovich, A.V., K.R. Jóhannsdóttir, A. Chella and B. Goertzel (eds.), *Biologically Inspired Cognitive Architectures 2010: Proceedings of the First Annual Meeting of the BICA Society*, Frontiers in Artificial Intelligence and Applications, 221, 195-244.

[8] Fodor, J.A. 1983, *The Modularity of Mind*, Cambridge, MA: MIT Press.

using a hybrid knowledge base that is able to process jointly different form of categorization and different kinds of categorized knowledge in form of complex structures of concepts: the DUAL PECCS.

The core of DUAL PECCS as a "cognitively inspired categorization system" (71) is a hybrid knowledge base, in which concepts are represented both according to the classical theory of concepts (a list of features of the concept itself, which are the necessary and sufficient conditions for a thing to be regarded as a member of the category expressed by the concept) and to the prototype/exemplar theories (using typical information about the concept):

> From a reasoning perspective, one of the main novelties introduced by DUAL PECCS consists of the fact that it is explicitly designed according to the flow of interaction between commonsense categorization processes (based on prototypes and exemplars and operating on conceptual spaces representations) and the standard rule-based deductive processes (operating on the ontological conceptual component) (73).

Conceptual spaces representation and ontologies are available and up-to-date tools to representing knowledge in an artificial system, so this can be considered an extension of cognitive architectures such as SOAR and ACT-R in their standard diagram but still in line with them. It is not surprising that the focus of the cognitive design approach is seen by the author in a development and an improvement of knowledge representation encompassing different theories of concepts to have a flexible behavior and performance in the artificial system from the point of view of knowledge. One of the main reasons of the birth of last decades approaches to AI has been the hard issues arisen by the "rigid" knowledge representation systems of AI in the 70s and 80s, and the general problem of how implementing common sense and background knowledge in an AI system, which cognitive architectures such as DUAL PECCS try, at least partially, to address. Lastly, even more interesting is the mention of a mutual influence of the implemented system and the experimental cognitive settings to which it is inspired, in the sense that the system performance can give some insights, in return, to the experimental research on the examined cognitive capability. According to the author, "this kind of result is exactly the type we look for in the context of a computationally grounded science of the mind" (75), and it is easily attributable also to the old and long-lasting tradition of the cognitive/psychological AI.

A last remark is needed about the notion of plausibility, as it is at the core of the modeling methodology in AI cognitive systems. The author stresses "the irrelevance, with respect to the 'plausibility' issue, of the level of abstraction adopted to model a given cognitive behaviour" (47). This position is somewhat controversial, as it is not approved by everyone. According to different approaches to cognitive modeling someone states that the right level of abstraction is the symbolic/logical/functional one, whereas others believe that the right level is the subsymbolical/neural/bodily one. The debate on such an issue has been foundational in AI and cognitive science development from an epistemological standpoint. Of course, it is related to the successful results of different approaches in modeling different cognitive capabilities along the wide range of what is meant to be cognitive. Lieto's proposal on plausibility—that is already claimed by

Cordeschi among others, as we said earlier—is deserving as an attempt to go beyond this debate and to treat every different approach with the same relevance, thus justifying hybrid artificial systems also from their structural point of view:

> the notions of both cognitive and biological plausibility, in the context of computational Cognitive Science and computational modelling, refer to the level of accuracy obtained by the realization of an artificial system, with respect to the corresponding natural mechanisms (and their interactions) they are assumed to model. In particular, cognitive and biological plausibility of an artificial system asks for the development of artificial models (i) that are consistent (from a cognitive or biological point of view) with the current state-of-the-art knowledge about the modelled phenomenon and (ii) that adequately represent (at different levels of abstractions) the actual mechanisms operating in the target natural system and determining a certain behaviour (47).

The question about what elements in the structure of the natural system give rise to the behavior to be modeled is very consequent from these statements and the most relevant one concerning the epistemic and explanatory value of the model. Starting from the list of criteria to characterize biologically plausible robotic models proposed by Webb (2001),[9] Lieto provides his own list (called Minimal Cognitive Grid) that is more synthetic also to catch a more neutral plausibility dimension in evaluating the explanatory power of a model and that is based upon three main issues: the ratio between functional and structural elements in designing a model, its potential generality, and the performance match requiring relevant features in the natural system behavior such as errors and execution time.

The Minimal Cognitive Grid together with a general discussion of evaluating methods of artificial systems (and many examples and proposals of future line of related research) is one of the two main innovative contributions of the book as a study on the philosophy of artificial intelligence and cognitive science. The other one is the renewed strength that is given to the view that consider AI, at least as a relevant research opportunity, in the wide and multifarious range of its approaches as a cognitive discipline in its fundamentals, methods, and goals.

University of Bologna FRANCESCO BIANCHINI

Conant, James and Chakraborty, Sanjit (eds.), *Engaging Putnam*.
Berlin: De Gruyter 2022, pp. viii + 372.

Hilary Putnam has surely been a thinker of the first magnitude in the last quarter of the 20th century, providing first-class contributions to many fields in philosophy. Such contributions belong to subdisciplines like philosophy of science, philosophy of language, philosophy of mind, philosophy of mathematics, logic, epistemology, and ethics. Putnam's work has been so influential in many debates in these areas because of his readiness to change his mind when faced with compelling arguments, whether from himself or from other thinkers. Along the way, he has displayed an outstanding collection of different views and ideas—and many

[9] Webb, B. 2001, "Can Robots Make Good Models of Biological Behaviour?", *Behavioral and Brain Sciences*, 24, 6, 1033-50, DOI: 10.1017/s0140525x01000127

versions thereof. This variety can be difficult to track for common readers and sometimes even for scholars.

The present collection, *Engaging Putnam*, edited by James Conant and Sanjit Chakraborty, is a major attempt to keep alive various relevant threads in Putnam's legacy and to honour an absolutely leading figure in contemporary philosophy. They are not shy to acknowledge the difficulties in an enterprise like this, with so many arguments and views changed within a few decades—a philosopher that has been considered a "moving target" (16-20). However, this ensemble of views is in an important way tied together by a central thread in Putnam's efforts, the issue of realism understood as our struggle to grasp the crucial role that a mind-independent reality plays in our intellectual endeavours. The book has two introductions—one devoted to celebrating Putnam's greatness and uniqueness in the contemporary scene, and another to present the contents of the collection—and twelve chapters by philosophers whose work has been heavily influenced by Putnam's. The list includes renowned figures such as Yemima Ben-Menahem, Tim Button, Roy Cook, Mario De Caro, Maximilian de Gaynesford, Gary Ebbs, Sanford C. Goldberg, Tim Maudlin, Martha C. Nussbaum, Duncan Pritchard, Joshua R. Thorpe, and Crispin Wright. Almost all the chapters address from a specialist's perspective some particular view or argument by Putnam. Hence, this is not just an honorary book: the authors celebrate Putnam's legacy by trying to engage with his views in a critical way. In this review there is not enough space to duly cover all the papers included. I extend my apologies for concentrating on the contributions that better fit my personal appreciation of Putnam's work and/or spare my limitations of competence.

I start with Thorpe and Wright's essay on a topic of great relevance for Putnam's role in recent philosophical discussions: the controversial proof for the view that we are not brains in a vat (BIV).[1] Thorpe and Wright engage in a commendable goal: to figure out the main lessons from this argument and the ensuing 35 years of worldwide discussion. This is a very important goal, since the significance of the proof has "remained stubbornly controversial" (63). Because of this fact, the authors raise important questions: "Does the proof work? If so, what exactly does it show? And of what, if any, significance, metaphysical or epistemological, is the result?" (63). They lay out the argument as follows: "(1) If you were in the VAT scenario, you could not refer to BIVs. However: (2) You can refer to BIVs (since, of course, your word "BIV" refers to BIVs). Therefore: (3) You are not in the VAT scenario" (65). They discuss it first at the level of reference (65-66) and declare that the proof here works by means of the semantic externalism defended in terms of the Twin-Earth thought experiment. However, they argue that the status of premise (2) remains controversial: is it not question-begging for the overall argument? "[D]on't you have to know that you are not in the VAT scenario before you can know that you can refer to BIVs—and thus know exactly the thing that the VAT argument is supposed to prove?" (66). Then they proceed to read the argument at the level of concepts (66-67). Here the argument goes as follows: "(1*) If you were in the VAT scenario you could not have any concept of a BIV. But: (2*) You do have a concept of a BIV. Therefore: (3*) You are not in the VAT scenario" (67). This version is also supported by semantic externalism, now concerning conceptual content, and works as much as the former does—with the

[1] Putnam, H. 1981, *Reason, Truth, History*, Cambridge: Cambridge University Press.

same doubts concerning the (question-begging) status of premise (2) of the referential version. They then directly address this controversy (68-88). First of all, they show that the argument shares problems with McKinsey's argument,[2] enabling a thinker to gain contingent socio-linguistic knowledge from the armchair—this paradoxical conclusion is taken as evidence that even though these arguments may be formally valid, they fail to transmit justification to their conclusions.[3] Second, given these problems with the warrant of transmission it follows that, even though we do not conclude that the proof has failed, we face another issue concerning what it is that the argument is supposed to prove—it seems that, except for a sense in which the VAT argument succeeds, it depends on the fact that a VAT could not make the argument because this presupposes an unavailable mastery of the English language and because BIVs fail to refer to BIVs in their VAT language. Third, according to the authors, the many new sceptical versions of the thought experiment fail in the end to make the VAT argument unsuccessful, even though answering the sceptic was not Putnam's primary goal.[4] Finally, the main goal of the VAT scenario was to illustrate how metaphysical realism was not incompatible with errors in the ideal theory and indeed with the conception of an Ideal Error—the authors here show how a problem of the VAT scenario is its inability to see alternative options like Davidson's[5] to this unwarranted conclusion, as these permit to highlight significant differences between "metaphysical realism, understood as throughout this discussion, and Ideal Error" (87-8).

Another chapter which delves into Putnam's ground-breaking work is the one written by Goldberg, addressing the compatibility of semantic externalism with our understanding of the first-person perspective (107-129). Goldberg characterises semantic externalism, both for linguistic meaning and for mental content, as the acceptance of the following principles:

> LE [Linguistic Externalism] For all languages L and speakers S of L, there are some expressions e of L for which the standing meaning of e as used by S does not supervene on S's bodily states (107).
>
> AE [Attitude Externalism] For all subjects of the propositional attitudes S, there are some attitudes A of S's which are such that the fact that S instantiates A does not supervene on the facts constituting S's bodily states (108).

Goldberg then addresses the second topic, which is the first-person perspective, i.e. our epistemic perspective on the world, by distinguishing two conceptions: a *spatial* view and an *informational* view. According to the spatial conception, "to have a point of view—an epistemic perspective on the world—is to occupy a particular spatial location at every moment at which one exists" (109). According to the informational conception, "to have a point of view […] is to be such that one's cognitive life can be represented as an ever-evolving stock of information resident

[2] McKinsey, M. 1991, "Anti-Individualism and Privileged Access", *Analysis*, 51, 1, 9-16.
[3] "If one specific kind of epistemic basis for the premises of a valid argument is such that it would be *undermined* by doubt about its conclusion, then one cannot rationally be open-minded about the status of that conclusion yet simultaneously avail oneself of that basis to accept the premises" (73).
[4] See also Pritchard's chapter on this issue (263-64).
[5] Davidson, D. 1986, "A Coherence Theory of Truth and Knowledge", in Lepore, E. (ed.), *Truth and Interpretations*, Oxford: Blackwell, 307-19.

"in" one's information-processing system" (109). These options are compatible with each other: we can admit that the information which we access and process depends on the locations we find ourselves in at certain given moments (109-10). Goldberg adds further assumptions to this scenario, like the following: "the informational system just is a physical system that traces a spatial position through time" (110); and "novel empirical information" reduces to what has "causal impact on the physical system" (110). By putting these assumptions together, we can claim that one's point of view can be understood in terms of the location occupied, the initial state of the system, and all "*the physical goings-on within that system*" concerning its "*impacts*" with the world (110). While this conception is *prima facie* reasonable, it has a problem with AE: this picture of a first-person perspective only concerns causal relevance, while AE acknowledges the relevance of objects/other subjects in one's environment to characterise metaphysically one's mental life. According to Goldberg, this observation is the starting point of one greater difficulty, because AE challenges the usual conception of the autonomous epistemic subject (110-11). AE puts constraints on one's mental life: the concepts that form the contents of our attitudes cannot be specified independently of the subject's environment (111). Goldberg here affirms that many of us are tempted to say that there are dimensions of our mental lives that somehow escape AE's constraints (111). For example, whereas concepts are determined according to externalist credentials, "conceptions" may be more subjective, i.e. they can contain errors and idiosyncrasies, generating contexts which evade strict externalism. Goldberg reads Putnam's externalism as understanding this subjectivism as mostly wrong: conceiving of things cannot be specified independently of the world and the community a subject belongs to. But this puts the very idea of the autonomous epistemic subject in jeopardy (111). A new feature that may be useful and "tempting" in thinking about points of view is the idea that one's epistemic perspective on the world is metaphysically (though not causally) independent of the world itself (MIPOV). MIPOV seems plausible from the angle of introspection, that is, regarding "the nature of one's self-knowledge of […] the materials that constitute […] one's attitudes" (112), and gains traction also from considerations revolving around the idea of a conception. Without enough clues about how "conceiving" works, we would fail to capture how one takes the world to be (112). A problem is that such conceiving relies on a capacity to discern the content-relevant features of one's mental life from the armchair (112). But if this is the case, AE fails to plausibly account for the subject's point of view. Goldberg identifies the considerations concerning introspection as the main rationale for this conclusion after discussing the argument for it (114-15). As said, MIPOV exploits the concept of a "conception": an epistemic "perspective" on the world is captured by how one "takes things to be" (115). Goldberg argues that the level of conceptions is a level of description of the subject's mind that is metaphysically independent of how things are (116). This depends on an argument that exploits our ability to "hold the appearances fixed" while "varying the underlying reality" (116). At least in these circumstances, how a subject conceives of reality is metaphysically independent of that reality: "S's point of view can be invariant over how things are in the world; so any construal of her point of view that fails to appreciate this is deficient" (119). AE fails to appreciate how the construction of a point of view entails the ability to keep appearances fixed in the face of variations in how things are. Goldberg presents this argument as the crucial case against externalism in current debates. However, according to Goldberg, Putnam

already offered reasons to refute MIPOV, and so there is a challenge for the anti-externalist. Discussions revolving around introspection vs. AE have shown how externalism is compatible with discerning the commitments involved in representing things in a certain way from the armchair (121-22). Goldberg offers an analogous move for MIPOV's defence based on the contrast concepts/conceptions:

> Even in the restricted set of cases in which a subject accepts or presupposes that how things seem to her is indicative of how they are, how things seem to her—how they appear to her to be—can be held fixed, even as we radically vary the nature of the world around her (123).

Goldberg argues that we aim to represent objective kinds "as the objective kinds that they are" and this is a claim that can be endorsed even by Putnam's critics. This becomes the basis of an argument showing that "for any concept whose individuation is 'externalist', the subject's conception of that concept must be construed externalistically as well" (124).

Nussbaum's chapter addresses Putnam's relationship with Aristotle's legacy, dealing with some anti-reductionist lessons that became important in Putnam's later years.[6] The first lesson concerns the philosophy of mind and the way in which Putnam abandoned functionalism about mental states—i.e. the idea that mental states are identified in terms of the functional role they play in someone's cognitive economy. This ground-breaking idea permitted us to understand "abstract" computations as connected with a "material" substrate in a way inspired by the relationship between software and hardware. Nussbaum reconstructs how Aristotle's influence had a role in this important change of mind: it was in an Aristotelian spirit that Putnam at a certain point came to realise that the intentional level of mental states could not be reduced to the computational level required by machine functionalism. According to Putnam, the complexity of certain intentional states cannot be wholly explained in terms of computations, leaving aside the relations of such states with (sets of) objects in the real world (237). Another lesson with a distinguished Aristotelian flavour, according to Nussbaum, concerns the directional intentionality of thought and language. Putnam stated the superiority of Aristotle over Wittgenstein as a guide to this problem (238). Putnam started to wonder how Aristotle's idea of an isomorphic resemblance between the form of an object and the relative idea in one's mind anticipates a central insight of Wittgenstein's Tractarian picture theory of meaning. But these resemblances do not go too far: causal theories of reference put such insights quickly in jeopardy. Here, the idea of "not logically equivalent different descriptions of the same event" enters the scene. Therefore, the causal connection exploited by the causal theory of reference is alone insufficient and lacks an account of form (e.g. given Putnam's model-theoretic argument).[7] At this point, Aristotle and Wittgenstein take again the centre stage as both defend a particular notion of "form" (238). Putnam finds Aristotle's notion by far more useful than Wittgenstein's in dealing with the dispute with the causal theorist of reference. This choice is based on the worldly roots of Aristotelian metaphysics (while Wittgenstein's notion of form is abstract): according to Putnam, "[t]he idea that logic could do

[6] Also Ben-Menahem's chapter addresses the issue of reductionism (289-308).
[7] Putnam, H. 1981, *Reason, Truth, History*, Cambridge: Cambridge University Press.

all the work of metaphysics was a *magnificent* fantasy, but fantasy it surely was".[8] Another superior aspect of Aristotle's notion is its everyday (i.e. non-technical) character. An account like this is, however, exposed to objections. The first goes like this: our everyday representations sometimes go badly wrong, so these should not be inserted as criteria "in the mind in order for reference to be secured" (239). Putnam replied that the requirement of having the essential metaphysical properties always embedded in our everyday representations is too strict for getting reference right: "[p]eople successfully referred to water without knowing its atomic structure" (239). Another problem was Aristotle's idea that species have timeless essences, which is at odds with current biology. To this observation, Putnam replied by pointing out that even if timeless essences are hard to defend, certain features of them, such as "the ordinary synchronic notion of species" are still useful and indeed "indispensable" (239). Scholars now certify that Aristotle was not as rigid in defending "timeless essences" as medieval interpretations stated. Nussbaum concludes with another lesson concerning ethics that leaves also room for hints of Putnam's personality, providing a remarkable portrait (242-48).

The above chapters are just some highlights which can give the reader an approximate idea of what a great book this is. All the chapters would have deserved a full presentation as they tackle pivotal problems such as the a priori in philosophy of science, realism in philosophy of mathematics, scepticism in epistemology, free will, and naturalism, the ethical value of literature, and many more. This collection of papers on Putnam's work honours him by paying tribute to the central issues of his philosophy, without dodging going deep into the most controversial arguments, and often ending up with overt criticisms or noteworthy disagreements.

University of Cagliari PIETRO SALIS

[8] Putnam, H. 1995, *Words and Life*, Cambridge, MA: Harvard University Press, 71.

Advisory Board

SIFA former Presidents

Eugenio Lecaldano (Roma "La Sapienza"), Paolo Parrini (University of Firenze), Diego Marconi (University of Torino), Rosaria Egidi (Roma Tre University), Eva Picardi (University of Bologna), Carlo Penco (University of Genova), Michele Di Francesco (IUSS), Andrea Bottani (University of Bergamo), Pierdaniele Giaretta (University of Padova), Mario De Caro (Roma Tre University), Simone Gozzano (University of L'Aquila), Carla Bagnoli (University of Modena and Reggio Emilia), Elisabetta Galeotti (University of Piemonte Orientale), Massimo Dell'Utri (University of Sassari), Cristina Meini (University of Piemonte Orientale)

SIFA charter members

Luigi Ferrajoli (Roma Tre University), Paolo Leonardi (University of Bologna), Marco Santambrogio (University of Parma), Vittorio Villa (University of Palermo), Gaetano Carcaterra (Roma "La Sapienza")

Robert Audi (University of Notre Dame), Michael Beaney (University of York), Akeel Bilgrami (Columbia University), Manuel Garcia-Carpintero (University of Barcelona), José Diez (University of Barcelona), Pascal Engel (EHESS Paris and University of Geneva), Susan Feagin (Temple University), Pieranna Garavaso (University of Minnesota, Morris), Christopher Hill (Brown University), Carl Hoefer (University of Barcelona), Paul Horwich (New York University), Christopher Hughes (King's College London), Pierre Jacob (Institut Jean Nicod), Kevin Mulligan (University of Genève), Gabriella Pigozzi (Université Paris-Dauphine), Stefano Predelli (University of Nottingham), François Recanati (Institut Jean Nicod), Connie Rosati (University of Arizona), Sarah Sawyer (University of Sussex), Frederick Schauer (University of Virginia), Mark Textor (King's College London), Achille Varzi (Columbia University), Wojciech Żełaniec (University of Gdańsk)